HonestBait: Headline Generation via Faithful Forward Reference

Anonymous ACL submission

Abstract

Current methods for generating attractive headlines often learn directly from data, which bases attractiveness on the number of user clicks and views. Although clicks or views do reflect user interest, they can fail to reveal how much interest is raised by the writing style and how much is caused by the event or topic itself. Also, such approaches can lead to harmful inventions by over-exaggerating the content, aggravating the spread of false information. In this work, we propose HonestBait, a novel framework for solving these issues from another aspect: generating headlines using forward references (FRs), a writing technique often used in clickbait. A self-verification process is also included in training to avoid harmful inventions. We start with a preliminary user study to understand how FRs affect user interest, after which we present PANCO, an innovative dataset containing pairs of fake news with verified news for attractive but faithful news headline generation. Automatic metrics and human evaluations show that our framework yields more attractive results while maintaining high veracity.

1 Introduction

Fake news has become a medium by which to spread misinformation (Oshikawa et al., 2020; Vicario et al., 2019). One common way to fight against fake news is to release verified news. However, as the goal of news verification is to correct misinformation, their headlines are often bland, making it difficult to gain the attention of users, which works against the need to alleviate the harmful impact of fake news. Therefore, headlines for verified news articles should be rewritten to be more sensational but still faithful, which is expected to pique reader interest in verified news. Many studies have been conducted on generating styled headlines (Jin et al., 2020; Xu et al., 2019), among which *clickbait* represents the style that generates the most reads or clicks. Despite their success in attracting readers, there are several challenges in current clickbait generation models. First, clickbait datasets for training headline generators with sensational style transfer are commonly collected based on the amount of views or clicks, which assumes that headline popularity is always due to the writing style. However, user reading preferences could also be motivated by trending topics or major events. Although such headlines get many views and clicks, they could end up as noise in the dataset, harming sensational headline classification performance (Xu et al., 2019). Second, harmful hallucinations created by headlines exaggerated to be more sensational could distort the meaning of the original article. This is especially critical as we do not want our generation model itself to spread misinformation. However, as such sensational headline generation models often generate clickbait which contains more ambiguous words than a general abstractive headline, it increases the difficulty of evaluating faithfulness by aligning title semantics with the content of the article.

In this work, we propose making real news sensational by learning what fake news is good at. Quantity-wise, the many fake news articles can serve as learning materials by which to generate more attractive headlines; style-wise, fake news is written to attract attention. To learn these attractive writing styles, we adopt the forward-reference (FR) writing technique (Blom and Hansen, 2015), which draws from psychology and journalism, and is frequently used to create attractive headlines. Specifically, FR creates an information gap between users and the news article with the headline, which motivates user curiosity (Loewenstein, 1994) to investigate the news content, and hence provokes the desire to click on the headline. One example is the headline *"Wanna be an enviable couple? 12 things a happy couple must do... It's that simple!"*, which

drives readers to find out what those things are.

Here, to understand the relation between veracity, attractiveness, and FR types in news headlines, we conduct a preliminary user study to investigate the attractiveness of fake and real news, followed by an analysis of the FR types used in the selected headlines. Given these results and observations, we propose HonestBait, a novel framework by which to generate attractive but faithful headlines. In this framework, we use FR to remove the need to learn directly from the clicks-based dataset. To ensure the faithfulness of the generated headlines, we design a lexical-bias-robust textual entailment component on the generated headline and its original content to confirm that the latter infers the former. In addition, we propose PANCO, an innovative dataset which consists of pairs of fake and verified news headlines, their content, and their FR types. We conduct experiments on PANCO and evaluate the results by both autometrics and human evaluation. The contributions of our work are threefold:

- We conduct a thorough user study to understand the relation between reading preference and FR types on fake news and verified news.
- We propose a novel framework for generating attractive but faithful headlines. In human evaluations, HonestBait outperforms baselines on attractiveness and faithfulness.
- We propose a new dataset containing pairs of fake and verified news, including their titles, article content, and title FR types.

2 Related Work

Forward Referencing as a Lure Loewenstein (1994) shows how the desire for information motivates human curiosity. Forward-reference was later defined as a technique for creating curiosity gaps at a discourse level for use in headlines (Blom and Hansen, 2015; Yang, 2011). A similar concept is cataphora, in which information is forwarded as a teaser at a sentence level (Baicchi, 2004; Halliday and Hasan, 1976). Kuiken et al. (2017) investigate how editors rewrite headlines for digital platforms, and analyze the linguistic features of what makes up an attractive headline. Zhang et al. (2018) address attractive headline generation as question headline generation (QHG), which assumes that interrogative sentences are more popular. Indeed, such modality is a type of FR, but we argue that the interrogative style may not be suitable for some news headlines, especially for verified news.

Headline Generation Headline generation can be viewed as a more specific summarization task. Qi et al. (2020) propose a Transformer-based, self-supervised n-gram prediction objective. Liu (2019) propose BERTSum, a variation of BERT (Devlin et al., 2019) for extractive summarization. See et al. (2017) propose an attention-based pointer generator with a copy mechanism, which has made great progress in summarization. Although its capability of copying text from the source context is powerful, using it directly for verified news often leads to bland titles. Hence we apply forward references and a sensationalism scorer to produce more satisfying results. Xu et al. (2019) propose auto-tuned reinforcement learning to generate sensational headlines using a pre-trained sensationalism scorer, the resulting score of which is used as the reward to enhance the attractiveness.

Faithful Summarization Recent work investigates how to improve the faithfulness of the generated summary or headline. Matsumaru et al. (2020) propose pretraining a textual entailment scorer to filter out noisy samples in the dataset, preventing hallucination or unfaithful generation. Maynez et al. (2020) analyze the faithfulness of current abstractive summarization systems, and discover that textual entailment correlates better to faithfulness than standard metrics. Based on such work, one major direction is to evaluate generated summaries with textual entailment rather than raw metrics such as ROUGE (Lin, 2004) or BLEU (Papineni et al., 2002). Accordingly, we propose a faithfulness scorer based on textual entailment. During training, the scorer provides feedback in the form of a faithfulness score, which is used as an optimization goal in a reinforcement learning fashion.

3 Preliminary User Study

In this section, we investigate for a given topic which of the fake or real headlines users are more interested in, and how often forward references are found in interesting titles. Accordingly, we want to test the following two hypotheses:

H1: *Fake news headlines motivate user reading interest more than real news headlines.*

H2: *Forward references are commonly seen/used in headlines which interest users.*

We conducted the user study on both Chinese and English news to see whether forward refer-

ences were used across languages. For English headlines, we adopted FakeNewsNet (Shu et al., 2018), which contains fake and real news headlines about gossip and political news from GossipCop and PolitiFact. Since the real and fake news in FakeNewsNet are not paired up, we performed topical clustering to alleviate topical bias. For Chinese headlines, we directly leveraged news pairs labeled as *disagreed* in the WSDM fake news challenge dataset,[1] which contains one fake news headline and its corresponding verified news headline. In this way we avoid topical preferences.

We conducted the English user study using Amazon Mechanical Turk (Crowston, 2012). Each pair was labeled by 3 turkers, whereas each Chinese pair was annotated by 5 native speakers we recruited. To test H1, annotators chose which headline they wanted to read further, with four options: first headline, second headline, both and none. News veracity was not revealed during the study. Results show that both Chinese and English readers prefer fake news headlines. For Chinese, 39.75% of fake titles were chose, while 23.60% of real titles were chose. For English, the percentages are 34.57% and 30.33%, respectively. This result supports H1: fake news headlines motivate reading interest more than real news headlines.

To test H2, we asked another set of three annotators to label the FR type of the preferred headline selected in the previous user study. This question could have more than one answer as more than one FR type can be used to compose a title. The definition of each FR type are listed in appendix. Results show that 73% of Chinese and 85% of English headlines utilize FR techniques (at least one FR is included in the headline), which further supports H2: FR is commonly used in interesting headlines. For detailed distribution of reading preference and FR type of the preferred headlines, please see Appendix A and B.

4 Method

Having motivated the use of FR, we propose HonestBait, a novel framework which incorporates FR techniques. The left part of Figure 1 is a high-level workflow of the model. The model input during training contains verified news headlines and their content. First, we use an FR classifier to predict which FR type best fits the input real news headline, after which the generator takes the real news content as input and generates a headline. During each decoding step, we use the FR labels from the FR classifier to compute the FR type reward. After decoding, we utilize a faithfulness scorer and a sensationalism scorer to compute the faithfulness and sensationalism rewards by which to evaluate the generated headline. During inference, given verified news articles and their headlines, the framework generates interesting headlines using FR, and ensures fidelity via the faithfulness scorer. Thus the model has four major components: a sequence generator, a FR type classifier, a faithfulness scorer, and a sensationalism scorer. Below we describe these in detail.

4.1 Sequence Generator

We adopt a pointer network (See et al., 2017) as the sequence generator. For given real news content A with M tokens $\{w_1, w_2, \cdots w_M\}$ and its corresponding headline consisting of Q tokens $\{x_1, x_2 \cdots x_Q\}$, the encoder encodes each token with a bidirectional LSTM (Hochreiter and Schmidhuber, 1997) to obtain its hidden state h_t. We adopt Chinese word-level embeddings pre-trained on the Weibo corpus (Li et al., 2018). Following Luong et al. (2015), we use an attention mechanism on the hidden states of the encoder: $a_i = \text{softmax}(h_{ti})$, $h_e = \sum_i a_i h_{ti}$, and then h_e is sent to the decoder. For each token in the article, the decoder updates its hidden state as s_t. The probability distribution of each time step over the vocabulary is computed through two linear layers: $o_t = (\boldsymbol{W_v}[h_e \oplus s_t]) + b_v$, $P_{voc}(x^*) = \text{softmax}(\boldsymbol{W_v'} o_t + b_v')$, where $\boldsymbol{W_v}$, $\boldsymbol{W_v'}$, b_v, and b_v' are trainable parameters. The final distribution is combined with the probability computed by copy mechanism, making words from source article available for generation:

$$p_{gen} = \sigma(\boldsymbol{V_c^h} h_e + \boldsymbol{V_c^s} s_t + \boldsymbol{V_c^x} e_t + b_c),$$
$$P_{final}(x^*) = p_{gen} P_{voc}(x^*) + (1 - p_{gen}) \sum_i a_i, \quad (1)$$

where a_i is the attention weight over the input tokens computed by the encoder, and e_t is the embedding of the tokens in the article. $\boldsymbol{V_c^h}$, $\boldsymbol{V_c^s}$, $\boldsymbol{V_c^x}$, and b_c are trainable parameters, and σ is the sigmoid activation function. For the objective we use the negative log likelihood as $\mathcal{L}_{MLE}$ as follows:

$$\mathcal{L}_{MLE} = -\frac{1}{Q} \sum_i^Q \log P_{final}(x_i). \quad (2)$$

[1] https://www.kaggle.com/c/fake-news-pair-classification-challenge

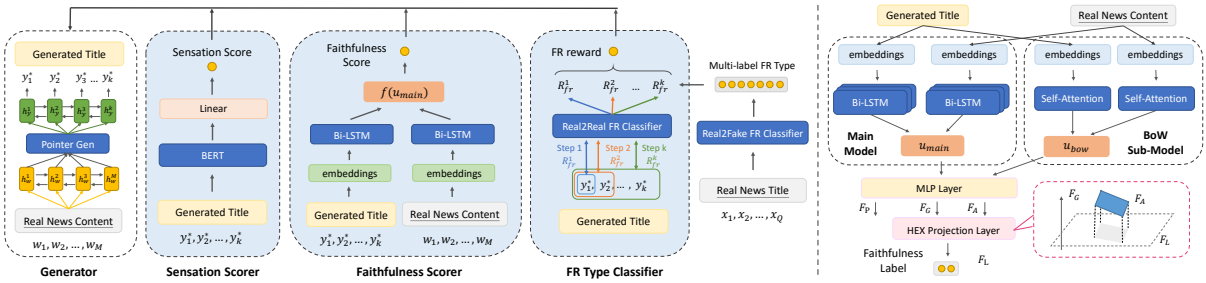


Figure 1: Left: The proposed framework. Required inputs are underlined. Components with frozen training weights are depicted with a blue background. Right: Workflow of model-level debiasing module.

4.2 Forward Reference Type Classifier

To make the model able to mimic different FR types, we pre-train two multi-label classifiers: (1) real2fake, which receives the real headline as input and predicts the FR type of the corresponding fake news; this classifier learns which kinds of real titles are suitable for which specific combination of FRs. (2) real2real, which predicts which FR type the generated headline contains; this is pre-trained by taking real news headlines as input and classifying which FR type these headlines have. We implement these FR classifiers with a BERT-based encoder. Given a verified news headline, we obtain a sentence-level representation h_v with the hidden state of the [CLS] token. The final prediction $\boldsymbol{p}_{fr} \in \mathbb{R}^l$ is the probability distribution of each forward reference type predicted by a MLP classifier following a sigmoid function; the FR type prediction is defined as:

$$\tilde{y}^i_{fr} = \begin{cases} 1, & \text{if } p^i_{fr} > \theta \\ 0, & \text{otherwise} \end{cases}, \quad \hat{y}_{fr} = [\tilde{y}^1_{fr}; \tilde{y}^2_{fr}; ...; \tilde{y}^l_{fr}] \tag{3}$$

where i denotes the i-th dimension of p_{fr} and θ is a threshold, here set to 0.5. $\hat{y}_{fr} \in \{0,1\}^l$ and l is the number of the FR type. We pre-train these models using binary cross entropy loss, yielding a 0.65 micro F1 score on real2fake and 0.91 on real2real on a pre-training test set. Below we denote real2fake prediction as $\hat{y}_f$ and real2real as $\hat{y}_r$, both of which are calculated using Eq. 3.

4.3 Forward Reference Reward

For each decoding time step, we calculate the FR reward once: tokens generated up to the current time step $y^*_{1:t}$ are sent to the real2real FR-classifier to derive $\widehat{y^{1:t}_r}$ and calculate how well the generated text fits the FR prediction by real2fake $\hat{y}_f$. This is formulated as:

$$R_{fr} = \frac{1}{T}\sum_{i}^{T}(1 - \text{MSE}(\hat{y}_f, \widehat{y^{1:i}_r})), \tag{4}$$

where MSE denotes the mean squared error and $R_{fr} \in [0,1]$ is the FR reward. In practice, we can also use the FR label of the fake news acquired from our user study y_{fr} to replace $\hat{y}_f$, and view this setting as an upper bound for real2fake accuracy to calculate R_{fr}. The FR prediction by real2fake $\hat{y}_f$ can be treated as a hint to guide HonestBait as to which FR type is more likely to be applied given the real news headline. Based on our observation, however, to interest the user, it is not necessary to generate headlines according to a specific FR type. Hence, real2fake can be used as an auxiliary tool to help decide which FR type to use, and is especially useful when the dataset contains no FR-type labels.

4.4 Faithfulness Scorer

Inspired by Kryscinski et al. (2020) that textual entailment better correlates to faithfulness than raw metrics, we use a pre-trained faithfulness scorer to evaluate whether the generated headline distorts or contradicts the corresponding content. When pre-training, we use a real news headline and its content as a positive example, and a fake news headline with the corresponding real news content as a negative example. We pre-train this as a natural language inference (NLI) task (classifying entailment and contradiction).[2] The sentence embeddings of headline and content are denoted as x_h and w_h. We apply a popular method to encode sentences for the NLI model (Conneau et al., 2017):

$$h = [x_h; w_h; x_h - w_h; x_h \odot w_h] \tag{5}$$

where ; denotes concatenation on the hidden dimension, and $\odot$ denotes the element-wise product.

[2]We exclude real-headline fake-content pairs since some of the fake contents are not necessarily fake.

Lexical Bias We discover that the verified news in the WSDM dataset often contains lexical bias; in particular, many true headlines entailed by their contents contain the words “verification" and “rumor" which results in a shortcut model, i.e., the NLI model tends to classify samples as entailment based simply on the existence of certain words. In addition, the word-overlap bias (WOB), i.e., high word-overlap (Naik et al., 2018), also harms our entailment task to ensure faithfulness. This is especially true when the fake headline in PANCO also has high word overlap with the verified content as they concern the same specific event or person. Thus, we adopt the model-level debiasing learning module (Zhou and Bansal, 2020) to our entailment task; its workflow is illustrated in the right part of Fig. 1. A bag-of-words sub-model is deployed to capture superficial features, since it has the least reasoning ability, and is more likely to use shortcuts to make predictions. The main model, in turn, consists of two bi-LSTM (Hochreiter and Schmidhuber, 1997) networks that are capable of reasoning over deeper semantics. During training, HEX projection (Wang et al., 2019) is used to screen out superficial features by making the hidden state of main model and the BoW sub-model orthogonal, allowing the main classifier to focus on deeper features. Following Zhou and Bansal (2020), we use a self-attention layer as the bag-of-words sub-model encoder, and a 3-layer bi-LSTM with skip connections and residuals as the main model encoder:

$$\begin{aligned}\mathbf{F}_A &= f([u_{bow}; u_{main}], \xi),\\ \mathbf{F}_P &= f([0; u_{main}], \xi),\\ \mathbf{F}_G &= f([u_{bow}; 0], \xi),\end{aligned} \tag{6}$$

where f denotes the final MLP classifier, u_{bow} and u_{main} denote the sentence embeddings encoded with Eq. 5, and ξ denotes the classifier parameters.

$$\mathbf{F}_L = (I - \mathbf{F}_G(\mathbf{F}_G^T\mathbf{F}_G)^{-1}\mathbf{F}_G^T)\mathbf{F}_A, \tag{7}$$

where $\mathbf{F}_L$ is projected to the orthogonal space of $\mathbf{F}_G$. We use the same training objective as Zhou and Bansal (2020). After training, the main model $\mathbf{F}_P$ is ready for use in the proposed framework as shown in Fig. 1 to make inferences and to calculate the faithfulness score $\mathcal{R}_{faith}$. The faithfulness scorer achieves 0.83 accuracy on the testing set.

4.5 Sensationalism Scorer

In addition to being faithful, the desired output must be attractive. We make use of another BERT-based binary classifier to obtain the sensationalism score. We first manually determine the news categories that are consistently sensational (fashion, gossip, headlines, international, society, politics) and collect the news headlines along with the content in these categories. The collected news headlines are regarded as sensational. For non-sensational headlines, we utilize a pointer generator to obtain a summary headline, and treat this as a non-sensational title since summarization models retain only the semantics of the content. We train the sensation scorer with binary cross entropy along with a softmax layer to produce a sensation score $\in [0, 1]$: $\mathcal{R}_{sen} = \sigma(\boldsymbol{W_s} x_s + b_s)$, where x_s is the aggregated representation of the [CLS] token produced by BERT, σ is the softmax function, and W_s and b_s are learnable weights. The accuracy on test set is 0.86, indicating its ability to discriminate sensational headlines.

4.6 Hybrid Training

We adopt the reinforcement learning (RL) algorithm (Williams, 1992) to train our model with the weighted sum of scores produced by the FR classifier, the faithfulness scorer, and the sensationalism scorer as the reward R. Following Xu et al. (2019); Ranzato et al. (2015), we use the baseline reward $\hat{R}_t$ to reduce variance, where $\hat{R}_t$ is the mean reward estimated by a linear layer for each time step t during training. The final reward and its loss are

$$\begin{aligned}\mathcal{R} &= \mathcal{R}_{fr} + \alpha\mathcal{R}_{faith} + (1-\alpha)\mathcal{R}_{sen}\\ \mathcal{L}_{RL} &= -\frac{1}{T}\sum_{i}^{T}(R - \hat{R}_t)\log P_{final}(x_t).\end{aligned} \tag{8}$$

Similar to Xu et al. (2019), we compute the final loss as the combination of $\mathcal{L}_{MLE}$ and $\mathcal{L}_{RL}$:

$$\mathcal{L} = \lambda\mathcal{L}_{MLE} + (1-\lambda)\mathcal{L}_{RL}, \tag{9}$$

where both λ and $\alpha \in [0, 1]$ are hyperparameters to balance the weight of each component, and the composite design here is to ensure that we produce headlines that satisfy all objectives. To sum up, we use the pre-trained faithfulness scorer to evaluate the textual entailment between the generated headline and the content, the sensationalism scorer to measure the sensationalism of the generated headline, and the FR type classifier to estimate whether the generated headline matches the given FR type.

5 Experiment

We conducted experiments to evaluate HonestBait. We describe the experimental dataset, following by the result of autometrics and human evaluation, and include a case study to further demonstrate the superiority of the proposed model. Source code is available at https://reurl.cc/q1DmyD

5.1 PANCO Dataset

We compiled **Pa**ired **N**ews with **Co**ntent (PANCO), a dataset that originated from a fake news classification competition held by WSDM. The competition involved a textual entailment task in which two news headlines were given as input: the task was to predict the relationship between the headlines. Each sample in the original dataset includes a fake news headline and a headline that is either *agreed* (two fake stories describing the same event), *unrelated* (two stories describing different events), or *disagreed* (two stories describing the same event, one of which is real and the other fake). We augmented the provided dataset in the following way: (1) We discarded *agreed* headline pairs (both titles are fake). (2) We manually examined the *unrelated* pairs and discarded unrelated samples because we found that some unrelated samples implicitly verify the fake headline. (3) We merged selected *unrelated* samples with *disagreed* samples. (4) We used each title as a query and used Google Search to determine the source of each news story. (5) We deployed a crawler to acquire news content from sources which matched the title. (6) Five annotators labeled the FR type of each headline; the final label was decided by majority vote. The proposed dataset has a total of 7,930 paired samples, containing fake news and the corresponding verified news along with their content and FR type. Some statistics are listed in Fig. 2. The main novelty of PANCO is the collection of pairs (describing the same event) of verified and fake news as well as the content. In addition, we provide the FR type label for each headline and for both real and fake news as linguistic features for further study. Also, although WSDM also provided English translations of the headlines, they were machine-translated and hence not satisfactory for further experiments; thus PANCO is a Chinese-only dataset.

5.2 Baseline and Settings

We compared the proposed model with the following strong baseline for Chinese headline generation.

Model	R_1	R_2	R_L	BS	FR
Ptr-G	41.86	28.18	37.30	69.61	55%
Clickbait	41.02	28.03	36.64	69.52	69%
ROUGE	43.75	27.65	35.65	71.56	59%
ProphetNet	**46.82**	30.40	38.89	**73.57**	50%
BertSum	28.09	16.15	18.86	63.22	17%
T5	44.27	28.55	38.66	72.73	60%
HonestBait	43.76	**31.45**	**40.42**	72.61	**80%**

Table 1: Automatic metric of proposed model against baselines. R_n represents the n-gram ROUGE score and R_L is the ROUGE-L score. BS represents the BERT score. FR is the ratio of generated headlines using FR.

The pointer-generator network (Ptr-G) (See et al., 2017) is an LSTM-based model with attention and a copy mechanism. Clickbait (Xu et al., 2019) uses a CNN-based sensationalism scorer to automatically balance MLE and reward loss, and also for use as a reward to generate more sensational headlines. ROUGE uses the Ptr-G architecture but with the ROUGE score as a reward. T5 (Raffel et al., 2020) is a text-to-text Transformer-based model; we utilize T5 with Pegasus (Zhang et al., 2020) pre-training to strengthen the baseline. ProphetNet (Qi et al., 2020, 2021) is a Transformer-based model that utilizes future n-gram prediction as a self-supervised objective. It performs strongly over several summarization benchmarks. For human evaluation and the case study, we also include Gold, which represents real human-written headlines as a strong baseline. Evaluation is done on PANCO by autometrics and human evaluation.

Experimental settings are detailed as follows. We first pre-trained all baselines on the LCSTS dataset (Hu et al., 2015) with 480,000 steps. LCSTS is a large-scale Chinese summarization dataset containing 2,400,591 samples with paired short text and summaries. We used the pre-trained weights to fine-tune all baselines on the PANCO training set for another 10,000 steps. The λ hyperparameter was set to 0.2 and α to 0.4. For qualitative analysis of λ and α, please refer to Appendix D and E.

5.3 Automatic Metrics

We used three autometrics for evaluation: ROUGE-n (Lin, 2004), ROUGE-L, and the BERT score (Zhang* et al., 2020). Though autometrics are in general not reliable for text generation (Sulem et al., 2018; Callison-Burch et al., 2006; Schluter, 2017), we still provide autometric results for reference. Results in Table 1 show the good abstractive ability of HonestBait with the highest 40.42 R_L score. Among the baselines,

		Headline	Content	Top-10 Words in Headline	Top-10 Words Eng. Trans.
Real	max	51	4731	辟谣, 谣言, 回应, 假,网传,	Clarify, Rumor, Response, Fake, Buzz,
	avg.	13.7 ± 3.8	576.8 ± 449.5	官方, 致癌, 吃, 造谣,真相	Official, Carcinogenic, Eat, Spread rumor, Truth
Fake	max	29	12653	吃, 网友, 致癌, 没, 女儿,	Eat, Netizens, Carcinogenic, Without, Daughter,
	avg.	13.9 ± 3.8	603.7 ± 566.9	喝酒, 酒, 驾, 开车, 算	Drinking, Wine, Drive, Driving Car, Belongs to

Figure 2: Headline and content length of PANCO dataset. We report the max and average length at the word level. We also list the top 10 most frequent words in real/fake headlines, sorted from left to right, top to bottom. Numbers and names are ignored.

ProphetNet is the strongest, with the highest R_1 and BERT scores, perhaps due to its n-stream self-attention mechanism. On the other hand, the extractive summarization model BERTSum performs worst here, as extracting a sentence from the article as its headline is not a common practice in general.

In the last column of Table 1, we further use the real2real FR classifier to detect which FR technique(s) the generated headlines are using, and report the percentage of generated headlines using FR in different models. The result shows that 80% of the headlines generated by HonestBait exploit FR to make headlines more attractive, which is the highest among all models, indicating that HonestBait indeed learns to utilize FR techniques during generation.

Model	ATRC	FAITH	FLCY
Ptr-G	-29.50	-17.83	-19.80
Clickbait	-6.00	-22.33	-9.25
ROUGE	-17.50	-17.25	-24.66
ProphetNet	-5.60	-5.50	4.33
BertSum	-30.50	-21.99	-9.70
T5	-12.50	-10.25	-1.25
Gold	-11.25	1.00	8.34
HonestBait	-	-	-

Table 2: Results of pairwise comparison, in terms of attractiveness (ATRC), faithfulness (FAITH), and fluency (FLCY), shown as percentages.

5.4 Human Evaluation

We conducted a human evaluation to further evaluate the attractiveness, faithfulness, and fluency of the generated headlines. We randomly selected 100 samples from the test data in PANCO, and asked 5 native speakers to select headlines in response to the following questions: (1) Which headline makes you want to read further? (2) Which headline is more faithful to the content? (3) Which headline is more fluent? The workers were given two generated titles and the story content, and were asked to select *first title*, *second title*, or *tie* in response to the questions. The order of titles was shuffled and the generation system behind was not revealed.

Table 2 reports the pairwise comparison results as percentages. Each number in the table is the competing model compared to the proposed HonestBait, following Zhao et al. (2020). For example, the output of Ptr-G is 12.50%/45.50%/42.00% better/same/worse than HonestBait in terms of attractiveness, resulting in 12.50%-42.00% = -29.50% in the table. Results show that for both attractiveness and faithfulness, HonestBait outperforms all baselines by a large margin. We believe this is due to the use of forward referencing and the faithfulness check. Compared to the pure click-driven attractiveness-optimized Clickbait (Xu et al., 2019), HonestBait outperforms by directly learning writing skills to avoid other impact factors of attractiveness. In addition, boosting only attractiveness makes Clickbait relatively unfaithful (-22.33). As for fluency, only ProphetNet and Gold outperform our model. As we did nothing specifically to improve fluency like ProphetNet's n-stream attention, this result indicates that HonestBait maintains reasonable fluency while increasing attractiveness and faithfulness. Note that compared to human-generated headlines (Gold), HonestBait generates more attractive headlines (+11.25%) with only a modest drop in faithfulness (-1.00%), which shows the effectiveness of HonestBait for rewriting real news headlines to promote the stories.

5.5 Ablation Study

To further investigate our framework, we conducted an ablation study. We compared each setting with the full framework using the evaluation protocol from Sec. 5.4 by pairwise comparison. Results are shown in Table 3. Clearly, there is an significant drop in attractiveness when we remove the sensation scorer (-19.50%) or FR type reward (-16.00%), which indicates that even with the sensation scorer, attractiveness still decreases without the help of the FR reward (see setting *w/o FR*). That is, the FR reward indeed helps the model to learn

attractive writing styles. On the other hand, removing the faithfulness scorer results in the largest decrease in faithfulness (−11.50%). This also shows our lexical-bias-robust faithfulness scorer prevents deviations in the generated headline. This attests the effectiveness of each component in the proposed model.

	ATRC	FAITH	FLCY
W/o sen	-19.50%	-4.00%	-9.75%
W/o faith	-4.00%	-11.50%	-6.75%
W/o FR	-16.00%	- 5.50%	-6.25%

Table 3: Ablation study result. "w/o" represents removing corresponding component from the full framework.

Article: 几年前，女版"乔布斯"伊丽莎白·霍姆斯凭借一项革命性的血液检测技术——依靠一滴血就能进行两百多项专业检，被众人追捧。然而，好景不长，她的谎言很快被戳破，跌落神坛。...专家表示临床中液体活检还没有形成金标准，还不能完全代替组织活检... Few years ago, "the woman version of Jobs" Elizabeth Holmes became popular by proposing a revolutionary blood testing technique: using a drop of blood to perform over two hundreds of professional testing. But not for long, her lies were revealed, and she had fallen from the peak. Expert said the liquid biopsy in clinic has not yet formed a gold standard, and cannot completely replace tissue biopsy.
Ptr-G: 辟谣！乔布斯，真正告诉你真相！ Rumors! Jobs, really tells you the truth!
Clickbait: "辟谣"乔布斯能治疗试剂盒？真相看这里！ "Rumors" Jobs can heal the reagent box? Here's the truth!
Clickbait+ROUGE: 辟谣！乔布斯能让人验癌吗？ Rumors! Can Jobs make people test for cancer?
BERTSum: 一滴血就能测癌？ A drop of blood can detect cancer?
T5: 一滴血检测癌症是谣言，如何才能防癌？ A drop of blood can detect cancer is a rumor, how can we do to prevent cancer?
ProphetNet: 辟谣：滴血检测不能确定癌症 Clarification: Drop blood test cannot determine cancer.
Gold: 辟谣！一滴血「检测癌症」的真相 Rumors! The truth of using one drop of blood to test cancer.
HonestBait: 一滴血就能发现癌症？专家辟谣：假的！ A drop of blood can detect cancer? Experts clarify: it's fake!

Figure 3: Generated examples from different models. For simplicity, we only present part of the article.

5.6 Case study

Figure 3 shows an example illustrating headlines generated by different models. Results show that Ptr-G, Clickbait, and ROUGE extract the name "Jobs" from the article, which is a powerful ability of the copy mechanism to alleviate the generation of unknown tokens. However, in terms of being headlines, these texts are less satisfying in that they are not understandable. BERTSum and T5 make mistakes by generating open questions without answering them, which could motivate user interest but is not faithful enough for verified news headlines. T5 focuses on the wrong point borrowed from other articles as this article is not about cancer prevention, which could be harmful. In contrast, HonestBait generates interrogative sentences to attract readers, but with an explicit clarification of the fake information, and is aligned to the content.

Full: 吃木瓜可丰胸？营养师辟谣：不靠谱 Papaya is helpful for breast enlargement? Nutritionist clarified: not reliable.
w/o Faith. Scorer: 吃木瓜可丰胸？营养师辟谣：吃木瓜可丰胸. Papaya is helpful for breast enlargement? Nutritionist clarified: Papaya is helpful for breast enlargement.

Table 4: Examples w/ and w/o faithfulness scorer

We also provide examples generated with and without the faithfulness scorer, as listed in Table 4 to better demonstrate the effectiveness of the faithfulness scorer. The headline generated without the faithfulness scorer appears quite fanciful, while with the proper guide of the faithfulness scorer, it successfully produces a true headline. These examples confirm that the proposed HonestBait has good writing techniques, yielding attractive and faithful headlines, which is indeed indispensable for news headline generation.

6 Conclusion

In this paper, we present HonestBait, a novel framework for generating faithful but interesting headlines. Moreover, we construct PANCO, a novel dataset that includes the title and content of pairs of fake and verified news, along with their forward reference type for further research. Our user study show that verified news headlines are relatively boring, and forward references are used in most headlines that readers like. Experiment results show that HonestBait outperforms all baselines in both automatic and human evaluations, which demonstrates its effectiveness on generating attractive but faithful headlines. We expect HonestBait to help us rewrite monotonous real news headlines to increase their exposure rate to help combat fake news.

7 Ethical Consideration

HonestBait is only designed to assist journalists as a reference to write more user-desired and faithful headlines for verified news. While being similar to clickbait or attractive headline generation systems, HonestBait also has the risk to be used by unwanted malicious users to generate sensational

headlines for fake news. Additionally, HonestBait may misjudge an offensive or unethical headline as user-desired headline. Our goal is to leverage the existence of fake news as a learning material to fight against misinformation, by encourage users to read verified news. This system focuses specially on the alignment between headline and its content, in order to be aware of the potential harm of misinformation generation. We are calling users not to abuse HonestBait to produce false information.

References

Annalisa Baicchi. 2004. *The Cataphoric Indexicality of Titles*, pages 17–38.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Kevin Crowston. 2012. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

M. A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Baotian Hu, Qingcai Chen, and Fangze Zhu. 2015. LCSTS: A large scale Chinese short text summarization dataset. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1967–1972, Lisbon, Portugal. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, Lisa Orii, and Peter Szolovits. 2020. Hooks in the headline: Learning to generate headlines with controlled styles. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5082–5093, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Jeffrey Kuiken, Anne Schuth, Martijn Spitters, and Maarten Marx. 2017. Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, 5(10):1300–1314.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization.

George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin*, 116:75–98.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Kazuki Matsumaru, Sho Takase, and Naoaki Okazaki. 2020. Improving truthfulness of headline generation. In *ACL*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. pages 1906–1919.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ray Oshikawa, Jing Qian, and William Yang Wang. 2020. A survey on natural language processing for fake news detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6086–6093, Marseille, France. European Language Resources Association.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Weizhen Qi, Yeyun Gong, Yu Yan, Can Xu, Bolun Yao, Bartuer Zhou, Biao Cheng, Daxin Jiang, Jiusheng Chen, Ruofei Zhang, et al. 2021. Prophetnet-x: Large-scale pre-training models for english, chinese, multi-lingual, dialog, and code generation. *arXiv preprint arXiv:2104.08006*.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-sequence pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2401–2410.

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):1–67.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Michela Del Vicario, Walter Quattrociocchi, Antonio Scala, and Fabiana Zollo. 2019. Polarization and fake news: Early warning of potential misinformation targets. *ACM Trans. Web*, 13(2).

Haohan Wang, Zexue He, Zachary L. Lipton, and Eric P. Xing. 2019. Learning robust representations by projecting superficial statistics out. In *International Conference on Learning Representations*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8(3–4):229–256*.

Peng Xu, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2019. Clickbait? sensational headline generation with auto-tuned reinforcement learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3065–3075, Hong Kong, China. Association for Computational Linguistics.

Youwen Yang. 2011. A cognitive interpretation of discourse deixis. *Theory and Practice in Language Studies*, 1:128–135.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, Huanhuan Cao, and Xueqi Cheng. 2018. Question headline generation for news articles. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 617–626.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

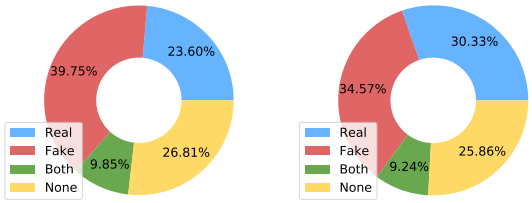
Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. pages 8759–8771.

Appendix

A Distribution of Reading Preferences

Reading preferences for Chinese and English are aligned in general. As mentioned in Sec. 3, For Chinese headlines, 39.75% of fake titles were chose more interesting than the real ones, while 23.60% of real titles win. For English headlines, the percentages are 34.57% and 30.33%, respectively. We report the complete distribution including tie situation as shown in Fig. 4, the ratio of tie situation is 36.66% and 35.10% respectively.



(a) Chinese headlines (b) English headlines

Figure 4: Reading preferences w.r.t. real news and fake news.

B Distribution of Forward Reference

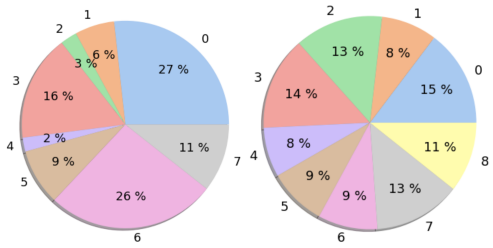
Here we report the distribution of FR type labelled by the annotators, as shown in Fig. 5. Table 5 summarizes the types of forward references, and their sample headline can be found in Fig. 6.

	Forward-referring Expressions
Type 0	None of below
Type 1	Demonstrative pronouns
Type 2	Personal pronouns
Type 3	Definite articles
Type 4	Ellipses
Type 5	Imperatives
Type 6	Interrogatives
Type 7	General nouns
Type 8	Location Adverbs

Table 5: Types of different forward-referring features

Note that for Chinese headlines, we merged type 8 into type 1 as their definitions are similar in Chinese.[3] The FR type used depends on the language or culture. We can easily find that the distribution of forward-referring features in Chinese headlines are more uneven. One possible reason is that the source of news in our dataset are less diverse, so the writing style is more monotonic. A majority part of fake titles utilize interrogatives (Type 6) to lure readers to look inside the article to search for the answer, while personal pronouns (Type 2) are less common to appear. As for English headlines, all types are observed roughly equally.

[3]In Chinese, location adverbs are often regarded as a type of demonstrative pronoun. Also, as very few samples are of type 8, we treat it as a special case of type 1.



(a) Chinese headlines (b) English headlines

Figure 5: Distribution of forward references in Chinese and English headlines with 2,800/2,159 samples. Type 0 indicates headlines without forward references.

C Stress Test of Faithfulness Scorer

To evaluate the robustness of the proposed faithfulness scorer to lexical bias, we conducted a stress test (Naik et al., 2018) by choosing samples (a) with high word overlap between headline and content but that are unfaithful, and (b) with headlines that contain certain cue words but are unfaithful. These two types of samples are biased but unfaithful, which is the hardest part for a faithfulness scorer. We selected (a) by taking fake news headlines with >85% word overlap with the verified contents. We then picked four word cues that contains the word "rumor". which we randomly added before or after headlines randomly chosen from LCSTS (Hu et al., 2015), after which we randomly selected other content from PANCO to pair with the headlines to form (b). In total, we collected 80 samples for (a) and 200 samples for (b). Note that headlines or contents selected from PANCO are in the validation and test sets. Accuracy results are 0.78 on (a) and 0.85 on (b), indicating an ability to reason semantically instead of taking shortcuts to determine faithfulness. Additionally, when we apply a biased faithfulness scorer (faithfulness scorer without debiasing) on HonestBait, we discovered that the only things it generates are "clarification" and "rumor", which shows the importance of debiasing mechanism in this framework.

D Analysis of different λ

Here we provide a qualitative analysis to examine the sensitivity of λ, which balances the weight of MLE loss and RL loss. In a sense, a higher λ leads to robust yet boring generation, as a higher λ relies more on MLE, and the MLE loss is calculated according to the gold title. Table 6 summarizes the generation with different λ. Note that the $\lambda = 0.0$ case is ignored, as it completely relies on RL loss, which often leads to broken generation results and is not practical in general. When $\lambda = 1.0$, the model completely relies on MLE loss and is identical to using only Ptr-G.

A smaller λ creates more diversity, and $\lambda = 0.2$ effectively balances the diversity, attractiveness, and fluency. Also, in $\lambda = 0.2$ and $\lambda = 0.6$, more sensational or eye catching words are used (highlighted in orange in Table 6), whereas $\lambda = 1.0$ shows a more plain, ordinary tone. When $\lambda = 1.0$, the generated results are unrelated and unintelligible, which also shows that our faithfulness scorer helps align the headlines to the content.

λ	**Generated Example**
0.2	临猗苹果滞销？山西临猗县政府辟谣：夸大失实 Apples from Linyi county are unsalable? Shanxi Linyi county government clarified: over-exaggerating.
0.6	临猗苹果滞销？事件是谣言！ Apples from Linyi county are unsalable? This event is a rumor!
1.0	临猗苹果滞销？电商客服：商家或涉嫌侵犯 肖像权Apples from Linyi county are unsalable? e-commerce's customer service: the merchant may violate portrait rights.

Table 6: Headlines generated with different λ. Orange words refer to more sensational expressions.

E Analysis of different α

We also conducted an analysis of how different values of α affect the generated headline. In Table 7, a lower α means more emphasis is put on sensationalism. A higher α tends to yield a relatively simple and monotonous sentence structure. In Table 7, we observe that when $\alpha = 1.0$, it predominantly generates affirmative sentences including "will" or "is", which are highlighted in red. On the other hand, a less dominant α provides more flexibility with respect to the sentence structure and adds diversity. When the reward is completely provided by the sensation scorer and the FR type reward ($\alpha = 0.0$), it seems that the model generates headlines from a different aspect and focuses on different keywords (highlighted in blue). However, such diversity comes at the risk of harmful invention, which we can see from the second dubious example. When $\alpha = 0.4$, the generated headlines maintain high veracity while improving attractiveness. Accordingly, we use 0.4 as our default setting. To better show these above-mentioned phenomena, we provide three examples here.

α	**Generated Example 1**
0.0	香菜吃多了不易于男性繁衍? Eating too much coriander is not good for male's reproduction?
0.4	谣言：吃香菜会杀精、阳痿、不易于男性 繁衍！Rumor: eating coriander will kill sperm, cause impotence and not good for male's reproduction.
0.8	谣言：香菜吃多了不易于男性繁衍！ Rumor: Eating too much coriander is not good for male's reproduction.
1.0	谣言：吃香菜会杀精、阳痿！ Rumor: Eating coriander will kill sperm and causes impotence.
α	**Generated Example 2**
0.0	辟谣：洗牙是没病防病的预防措施！ Clarification: Dental scaling is a precaution that prevents disease before it onset!
0.4	辟谣：洗牙能清洗牙齿？别再信了！ Clarification: Dental scaling can wash your teeth? Stop believing it!
0.8	辟谣：洗牙是一种保健！ Clarification: Dental scaling is a kind of health care!
1.0	辟谣：洗牙是清洗牙齿，是真是假！ Clarification: Dental scaling is washing your teeth, True or False!
α	**Generated Example 3**
0.0	辟谣！新衣服对身体会造成多大伤害？！ Clarified! How much harm will new clothes do to our body?!
0.4	甲醛致癌？是谣言！ Formaldehyde causes cancer? It's a rumor!
0.8	甲醛致癌？别再信了！ Formaldehyde causes cancer? Stop believing it!
1.0	甲醛致癌？是谣言！ Formaldehyde causes cancer? It's a rumor!

Table 7: Generated headlines with different α. Red words refer to monotonic affirmative, and blue words refer to more diversified expressions.

FR Type	Example Headline
1. Demonstrative pronouns	这是今年最大的养生谣言！ **This is** the biggest regimen rumor in the year!
2. Personal pronouns	据说这是他最后一部参与表演的电影 It is said that this is the last movie **he** participated.
3. Define articles	辟谣：嚼口香糖不能防止蛀牙 Rumor: Chewing gum does not prevent tooth decay.
4. Ellipses	接听了陌生Facetime，结果... After answering the unfamiliar Facetime, it turns out...
5. Imperatives	我的天！雪糕二次冷冻会产生可溶性毒蛋白? OMG! Will second freezing of ice cream produce **soluble toxic protein?**
6. Interrogatives	微波炉加热食物会致癌吗? Does microwave heating food cause cancer?
7. General nouns	天再冷也不能吃辣的**5**种人，吃了等于慢性自杀！ **5 kinds of people** who can't eat spicy food no matter how cold it is, eating them is tantamount to chronic suicide!
8. Location Adverbs	经常吃方便面真的会致癌吗？正确的解释在这里 Does eating instant noodles often really cause cancer? **Here** is the correct explanation.

Figure 6: Example headlines using different types of forward reference defined in Blom and Hansen (2015), retrieved from PANCO

Real Title	**Real Content**
辟谣：夏季曝晒后的瓶装水致癌 Rumors: Water in a bottle is carcinogenic after the exposure of summer sun.	夏天放在车里的瓶装水会致癌？又是一条健康谣言... Bottle water is carcinogenic in the car when it's summer? That's another health rumor...
Fake Title	**Fake Content**
BBC紧急曝光：这种水喝一口，就会致癌！ BBC urgent disclosure: this kind of water can lead to cancer with a sip!	今天，这个有害健康、甚至会夺人性命的巨大隐患终于被曝光了！世界卫生组织通报：9成以上瓶装水有毒... Today, this huge worry that is harmful and even taking lives has finally revealed! WHO announce: over 90% of bottle water is poisoned...

Figure 7: A sample of paired news collected from the PANCO dataset. Real news and fake news describe the same story: the real news headline is less attractive, and the fake news headline is more sensational.