CS529

Detecting Incongruent News Headlines With Auxiliary Textual Information

Team - HPVS-DataMining2023

Shubham Shankar 190101107

Harsh Jaiswal 190101039

Vinayak Bharadwaj 190101105

Pranav Vinchurkar 190101064

PHASE I : Detecting incongruent news headlines with auxiliary textual information

Introduction

News articles can affect people in a lot of ways. These articles can modify the way of thinking of a person as they want. A lot of times a person goes through an article after reading the headline of it, and if the headline doesn't match the body of the article the person can be filled with false information. Many times the articles are published with unrelated headlines to increase the advertising revenue. These type malpractices can even waste the time of the reader, because a reader cannot decide if the body is related or not before reading the whole article.The specific problem we tackle is the headline incongruence problem, where a headline of a news article holds unrelated or distinct claims with the stories across its body text. This type of incongruity in news stories is a major characteristic of clickbait. Incongruent headlines not only make a wrong impression on readers, but also become worse when shared on social media, where most users just share it without reading its actual contents. Therefore, it is crucial to develop automated approaches that detect incongruent headlines in news articles. Therefore, it is important to detect an incongruent news headline and help readers to read articles correctly, avoiding misinformation. Incongruent

news headline detection can apply to various areas such as clickbait and stance detection, however the lack of a large-scale dataset makes it difficult to train deep learning models to learn the complex relationship between a headline and its body, which is crucial in this problem.

Challenges in respective of research problem :

Incongruent news headline problems should be tackled separately from other types of news headline problems, such as clickbait and sensationalism as the headlines do not represent the information contained in the body accurately.The main challenge in this field of study is a lack of large-scale realistic datasets and many attempts have been made to tackle the problem such as The Fake News Challenge 2017 (FNC-2017) which provided 49,972 pairs of headlines and bodies with four labels: {agree, disagree, discusses,unrelated},and many others but most of the released datasets were restricted to title and body text, and the most of the methods only leveraged those features but there are many other textual information when people come across news articles specifically there are many other textual information when people
come across news articles the model proposed below consider these factors to be one of the important factors to detect incongruent news headline.

Model

Here, we define the news headline detection task as a binary classification problem, i.e., given a list of news article = $\{N\,1\,,\,\ldots\,,\,N\,K\}$, where K is the size of the dataset, each news article $N\,i$ is categorized as either congruent ($y\,i$ = 0) or incongruent ($y\,i$ = 1). Each piece of news article $N\,i$ is composed of a headline $H\,i$ and a $B\,i$, denoted as $N\,i$ = ($H\,i$, $B\,i$). The headline $H\,i$ consists of title $t\,i$ and subtitle $s\,i$, and the body $B\,i$ consists of body text $b\,i$ and image caption $c\,i$. Each of them are expressed as $H\,i$ = ($t\,i$, $s\,i$) and $B\,i$ = ($b\,i$, $c\,i$) respectively.
All the title, subtitle, and image caption are composed of l, m, and p words, respectively, and are expressed as follows:

$$t_i = \{w_1^{(t_i)}, w_2^{(t_i)}, \ldots, w_l^{(t_i)}\},$$

$$s_i = \{w_1^{(s_i)}, w_2^{(s_i)}, \ldots, w_m^{(s_i)}\},$$

$$c_i = \{w_1^{(c_i)}, w_2^{(c_i)}, \ldots, w_p^{(c_i)}\}.$$

The body text $b\ i$ has n sentences denoted as $b\ i$ = {bi1 , bi2 , … , bin}. Each sentence in the body text $b\ i$ is composed of max o words and is expressed as follows:

$$b_{ij} = \{w_{j1}^{(b_i)}, w_{j2}^{(b_i)}, \ldots, w_{jo}^{(b_i)}\}, \forall j.$$

1. **Word embedding layer**
   Each word w_q^r in the title, subtitle, body text, and image caption, is projected into d dimensional vector e_q^r by an embedding matrix $E \in R^{vxd}$ in an embedding layer, where v is the vocabulary size. We experimented with pre-trained fasttext embeddings to reflect postpositional attributes.

2. **Bidirectional word encoder**
   We adopt a bidirectional GRU to encode the information from the backward and forward words. Different GRUs are used to create meaningful representations for each textual information by encoding title,subtitle, body text, and image caption. Bidirectional GRU generates contextual representation h^r_q from word embedding e^r_q as follows:

   $$\overrightarrow{h}_q^r = \overrightarrow{GRU}^r(e_q^r),$$

   $$\overleftarrow{h}_q^r = \overleftarrow{GRU}^r(e_q^r),$$

   $$h_q^r = [\overrightarrow{h}_q^r; \overleftarrow{h}_q^r], \forall q, r,$$

   Where :

   $\overrightarrow{h}_q^r$ forward hidden states of GRU

   $\overleftarrow{h}_q^r$ backward hidden states of GRU

   $\overleftarrow{h}_q^r$ formed from concatenation.

3. Hierarchical body encoder

Inspired by the previous studies that encoded multiple sentences using hierarchical architecture, we used another bidirectional GRU to encode the body from the word level to sentence level. To guide our model to understand the overall representation of a sentence, we first fed the hidden states of both sentences in the body text and image caption to an average pooling layer to encode the word sequences to one dimensional vector. $e^{(\text{bi}_j)}$ is the pooled sentence vector in the body text and $E^{(\text{bi})}$ is the body text matrix that concatenate the sequence of n sentence vectors. $e^{(\text{ci})}$ is the pooled image caption vector .

$$e_j^{(b_i)} = \frac{1}{o}\sum_{t=1}^{o} h_{jt}^{(b_i)}, \forall j,$$

$$E^{(b_i)} = [e_1^{(b_i)}; , \ldots, ; e_n^{(b_i)}],$$

$$e^{(c_i)} = \frac{1}{p}\sum_{t=1}^{p} h_t^{(c_i)}.$$

To encode the complex textual relationship between the body text and image caption, we concatenated the caption vector $e^{(\text{ci})}$ to body text matrix $E^{(\text{bi})}$, then fed it to the bidirectional GRU to reflect backward and forward contextual information of the body as follows:

$$\overrightarrow{E}^{(B_i)} = \overrightarrow{GRU}^{body}([E^{(b_i)}; e^{(c_i)}]),$$

$$\overleftarrow{E}^{(B_i)} = \overleftarrow{GRU}^{body}([E^{(b_i)}; e^{(c_i)}]),$$

$$E^{(B_i)} = [\overrightarrow{E}^{(B_i)}; \overleftarrow{E}^{(B_i)}],$$

4. Subtitle to title attention

Similar to the body, we leveraged a subtitle, which is an additional textual information, to guide the model to understand the complex textual representation of the headline. As title and subtitle contain high density information in one or two sentences, we padded with maximum length to

let model reflect the meaning of all words. However, according to the dataset, 50% of the subtitle was missing value. Rather than using all the words in the subtitle, we obtained the overall representation via feeding the subtitle to average pooling layer as follows:

$$e^{(s_i)} = \frac{1}{m}\sum_{t=1}^{m} h_t^{(s_i)},$$

Where $h_t^{(si)}$ is the contextual representation of each word in the subtitle obtained in Section 4.2 and $e^{(si)}$ is a subtitle vector. To highlight the words in a title that contain crucial information about the headline, we used attention mechanism to attend to important words in the title with the subtitle. $A^{(ti)}$, expressed as Eq. (15), is the attended title highlights the meaningful word in the title.

$$A^{(t_i)} = \alpha_i e^{(s_i)^T},$$

where $\alpha\, i$ is an attention distribution and can be calculated by Eq. (16).

$$\alpha_i = softmax(E^{(t_i)^T} e^{(s_i)}). \tag{16}$$

In this equation, $E^{(ti)}$ = $[h_1^{(t_i)}; , \ldots , ; h_l^{(t_i)}]$ is the title matrix, which (t)concatenates the hidden states of the title and $h^{(t\,i)}$ is the contextual representation of each word in the title obtained in Section 4.2. We use all the title, attended title, and subtitle by concatenation to derive the headline representation of news article $E^{(hi)}$ as follows:

$$E^{(H_i)} = [E^{(t_i)}; A^{(t_i)}; e^{(s_i)}].$$

5. Attentive headline encoder

After obtaining headline representation $E^{(Hi)}$ and body representation $E^{(Bi)}$, we used attention mechanism to capture the relationship between the headline and body of the news article. $A^{(Hi)}$ is attended headline incorporating a body information and is expressed as

$$A^{(H_i)} = \beta_i E^{(B_i)^T},$$

where β_i is an attention distribution over headline and can be obtained by

$$\beta_i = softmax(E^{(H_i)^T} M^{(Ei)}).$$

Further, we concatenated original headline $E^{(Hi)}$ and attended headline $A^{(Hi)}$ to encode news article representation $E^{(Ni)}$ as follows:

$$E^{(N_i)} = [E^{(H_i)}; A^{(H_i)}].$$

6. Classifier

We then fed news article representation $E^{(Ni)}$ to the feed forward network to capture some features of news article representation. We use a 2-layer feed forward network with dropout in each layer, as shown in Fig. 3. The structure and hyperparameters of the feed forward network is obtained via grid search. Subsequently, we used a sigmoid layer to classify an incongruity of the news article N_i as follows:

$$\hat{p}_i = \sigma(af(E^{(N_i)}) + b), \tag{21}$$

where $f(\cdot)$ denotes the feed forward network, σ is a sigmoid function, a and b is a weight and a bias term in sigmoid layer, and p_i is a predicted probability of incongruity of the news article N_i. Then, the model is trained to minimize the following loss

$$L = -\frac{1}{K}\sum_{i=1}^{K}[y_i log(\hat{p}_i) + (1 - y_i) log(1 - \hat{p}_i)], \tag{22}$$

where L is a binary cross entropy function and y_i is an incongruent label of news article N_i .

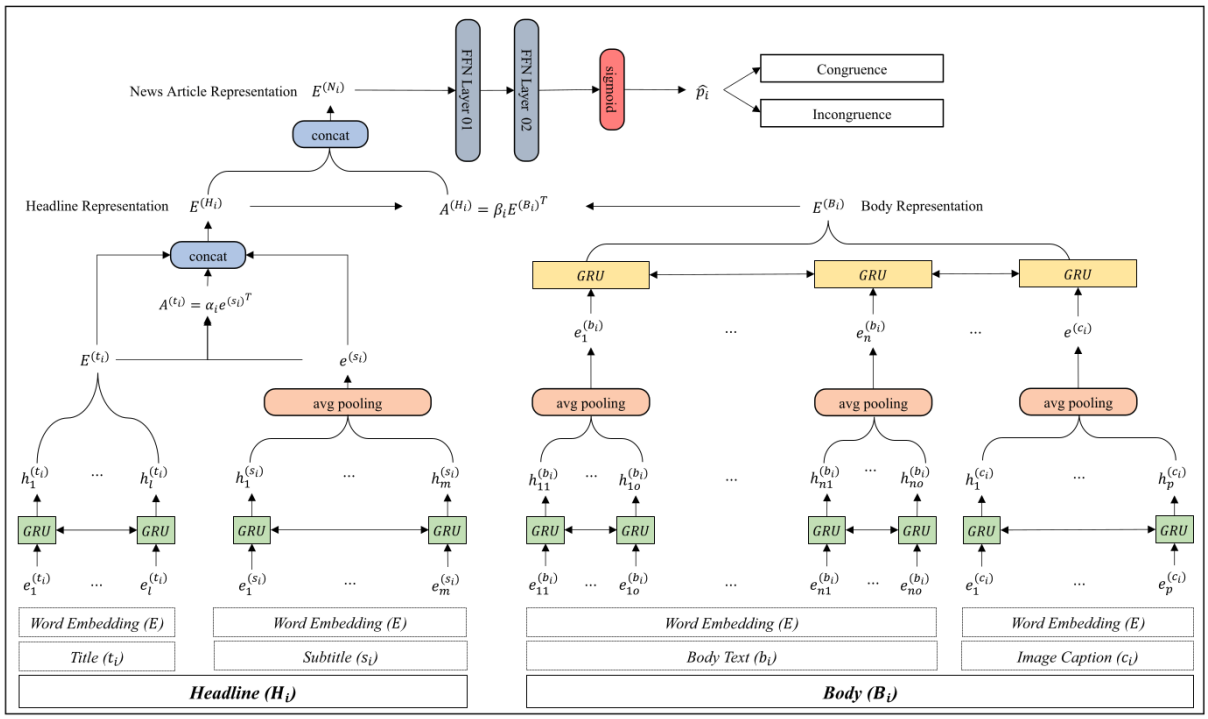

Fig. 3. Framework of the proposed model.

PHASE II : Topic Modeling

Limitation of Proposed Models

Determining the topics present in the headline and the body can play a vital role in detecting incroguency. The current model however does not perform any kind of topic detection.

Objective

The primary objective in phase 2 was to improve the model by considering the topics included in the body and the headline for each news article in the dataset.

Intuition

1. We observed that in multiple news articles where the headline and the body were incongruent, the topics in the headline were not included in the body.
2. In the news articles where the body and the headline were congruent, both the body and the headline addressed similar topics.
3. These observations led us to conclude that the topics included in the body and the headline can be an important factor in detecting incongruence.

Experiment Setup

1. **Topic Modeling**
    a. BERTopic was used to identify the topics present in the text corpus.
    b. After applying BERTopic , the topics and the top 20 words that represented each topic were generated as output.
    c. For the model we only considered the top 300 topics ranked on the basis of size (that is the number of documents belonging to a particular topic).
    d. For each word in the text corpus, the word's importance towards each of the 300 topics was stored in the form of a matrix. For each word a 300 dimension vector was obtained indicating the word’s

importance to each of 300 topics that were previously generated. This information would later be used in the embedding section of our model.

2. Embedding

a. Pretrained glove embedding was used to get a 100 dimension vector for each word.
b. This vector was concatenated with the 300 dimension vector generated in the topic modeling section which would be finally used in the model.

3. Bidirectional LSTM

a. From each document 12 words were chosen from the headline and 300 words were chosen from the body. The word embeddings generated before were taken as inputs.
b. The word embeddings of the words present in the headline were passed through the bidirectional lstm and the word embeddings of the words present in the body were passed through a different bidirectional lstm. The outputs of both the lstm would be used in the next layer.

4. Pooling and Vector generation

a. Mean pooling was applied to the outputs of the bidirectional lstm mentioned earlier to get a single vector representing the body and another single vector representing the headline. These vectors were of the same dimension.
b. Let ha denote the headline vector and hb denote body vector. A new vector which was the concatenation of (ha - hb) , (ha.hb) , ha and hb was generated. This vector would be passed to the next layer.

5. Fully Layer Connected Layer

a. The output of the previous layer was passed through a fully connected layer which used ReLU as the activation function.

b. This would then be passed to the final output layer.

6. Output Layer

a. The output of the previous fully connected layer was passed through another fully connected layer, which used softmax as the activation function.
b. Cross Entropy was used as the loss function.

Result

The accuracy of the model in phase 2 using Topic Modelling was 0.72.
The accuracy of the model used in phase 1 was 0.70.

Observation

After incorporating topic modeling an increase in the accuracy was observed.