MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching

Group: Dora_code

214101019 - Aishwarya Gupta
214101026 - Kumar Sandip Roy
214101048 - Sanjeev Kumar
214101051 - Shambhavi Shanker

List Of Contents

1. Introduction

Digitalization has made the flow of information around the world faster and easier. However, with the increase in information the chances of it being deceptive also increases. One such type of common information are News articles. Readers usually tend to read news headlines and forward it without reading the content; this can be dangerous for society as news headlines may not be with respect to the news body or vice versa. News headlines that incorrectly represent the content of the news body are called incongruent or click-baits. Thus it has now become crucial to devise a system capable of identifying news headlines as congruent or incongruent.

2. Problem Definition

The objective of our project is to identify whether the given news headline and content is congruent or not.News headlines that incorrectly represent the content of the news body are called incongruent.

In sum, the major contributions of this work are as follows:

(1) We are the first to use inter-mutual attention based semantic matching to detect incongruent news headlines. The key idea is to get the pairwise difference between word embeddings of original and synthetic headlines and compute a mutual attention matrix after applying a dense layer. Subsequently, row-wise max-pooling can be used to compute the attention scores, to be used for the classification.

(2) We use synthetic headlines generated from various generative adversarial networks based schemes using news body content to get the effective, contextual, and low dimensional representation.

(3) We also investigate two additional variants of the proposed model, which incorporate addition and concatenation of word embeddings of the word pairs of original and synthetic headlines.

(4) We combine all the three variants of word embedding operations in a clubbed model, which outperforms the three variants individually.

(5) We conduct experiments with two publicly available datasets, which show the effectiveness of the proposed models to detect incongruent news headlines.

3. Application of Project

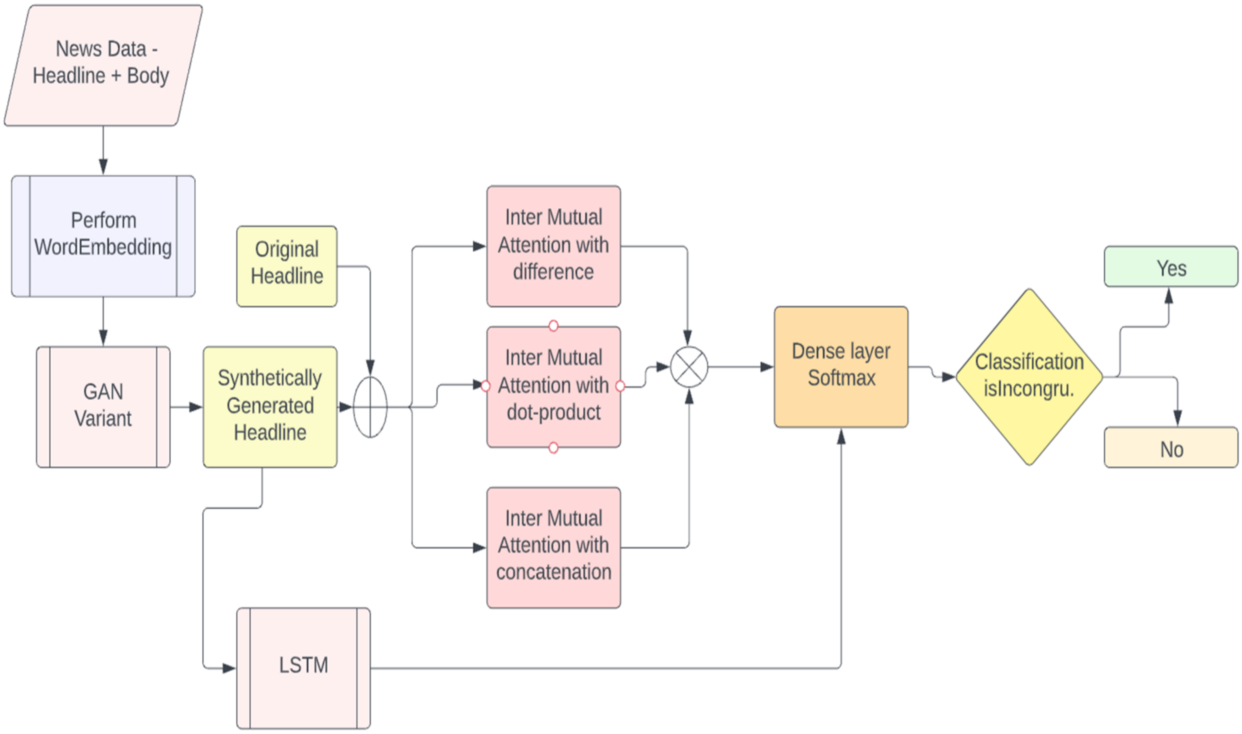
The objective is to come up with a system capable of identifying the news article as congruent or incongruent.This system can further be used in any news application to help stop the spread of such incongruent news.

4. Proposed Model

The paper”**MuSeM: Detecting Incongruent News Headlines using Mutual Attentive Semantic Matching**“ proposes a semantic matching technique based on inter-mutual attention that uses a synthetically generated headline corresponding to the news body content and original news headline to detect the incongruence.For generation of synthetic headline we have used the idea of text summarization as mentioned in “ **Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks**”.

5. Basic Working of our system

1. We will be first performing word embedding using a pretrained GLOVE model with dimension 300.
2. Further, we will generate synthetic headlines using a multitasking model consisting of GAN and Reconstructor.
3. We will perform intermutual attention of original and synthetic headlines by performing concatenation, dot-product, and difference.
4. We will use the softmax function to identify if news is congruent or not.

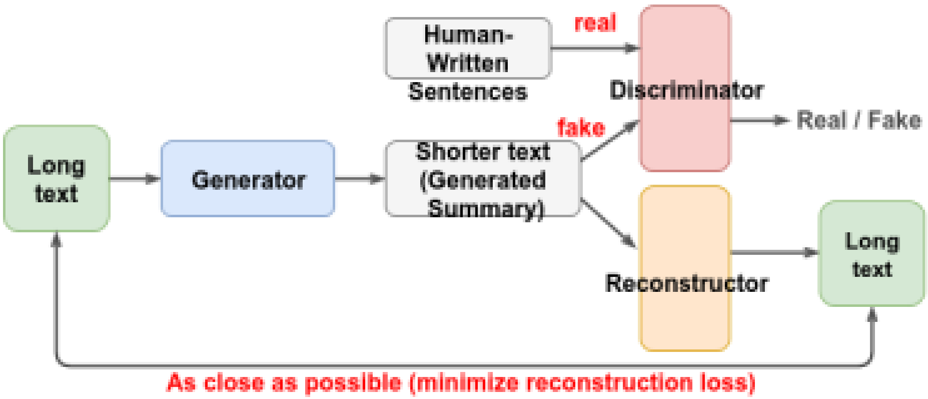

News Data - Headline + Body
Perform WordEmbedding
GAN Variant
Original Headline
Synthetically Generated Headline
LSTM
Inter Mutual Attention with difference
Inter Mutual Attention with dot-product
Inter Mutual Attention with concatenation
Dense layer Softmax
Classification isIncongru.
Yes
No

6. Synthetic Headline Generation

The model for synthetic Headline Generation is composed of three components:

1. generator G
2. Discriminator D
3. Reconstructor R.

Both G and R are seq2seq hybrids taking a word sequence as input and output a sequence of word distributions. Discriminator D, on the other hand,takes a sequence as input and outputs a scalar.



7. Discriminator1

D1 is a deep CNN with residual blocks, which takes a sequence of word distributions as input and outputs a score.The discriminator loss D_{loss} is

$$D_{loss} = \frac{1}{K}\sum_{k=1}^{K} D_1(G(x^{(k)})) - \frac{1}{K}\sum_{k=1}^{K} D_1(y^{real(k)}) \\ +\beta_1 \frac{1}{K}\sum_{k=1}^{K} (\Delta_{y^{i(k)}} D_1(y^{i(k)}) - 1)^2,$$

where K denotes the number of training examples in a batch, and k denotes the kth example. The last term is the gradient penalty.

8. Discriminator 2

The discriminator2 D2 is a unidirectional LSTM network which takes a discrete word sequence as input. At time step i, given input word y^s i it predicts the current score s_i based on the sequence $\{y^1, y^2, ..., y^i\}$. The score is viewed as the quality of the current sequence.

In order to compute the discriminator loss D_{loss}, we sum the scores $\{s_1, s_2, ..., s_N\}$ of the whole sequence y^s to yield

$$D_2(y^s) = \frac{1}{N}\sum_{n=1}^{N} s_n.$$

where N denotes the generated sequence length. Then, the loss of discriminator is

$$\begin{aligned} D_{loss} = \frac{1}{K}\sum_{k=1}^{K} D_2(y^{s(k)}) - \frac{1}{K}\sum_{k=1}^{K} D_2(y^{real(k)}) \\ +\beta_2 \frac{1}{K}\sum_{k=1}^{K} (\Delta_{y^{i(k)}} D_2(y^{i(k)}) - 1)^2, \end{aligned}$$

Using losses from D1 and D2, the training is being done to generate better summarization of the news content.

9. Self-Critical Generator

Since we feed a discrete sequence y_s to the discriminator, the gradient from the discriminator cannot directly back-propagate to the generator due to the vanishing gradient problem. Here, we use the policy gradient method. At timestep i, we use the i − 1 timestep score s_{i-1} from the discriminator as its self-critical baseline. The reward r D i evaluates whether the quality of sequence in timestep i is better or worse than that in timestep i − 1. The generator reward r D i from D2 is

$$r_i^D = \begin{cases} s_i & \text{if i = 1} \\ s_i - s_{i-1} & \text{otherwise.} \end{cases}$$

However, some sentences may be judged as bad sentences at the previous timestep, but at later timesteps judged as good sentences, and vice versa. Hence we use the discounted expected reward d with discount factor γ to calculate the discounted reward d_i at time step i as

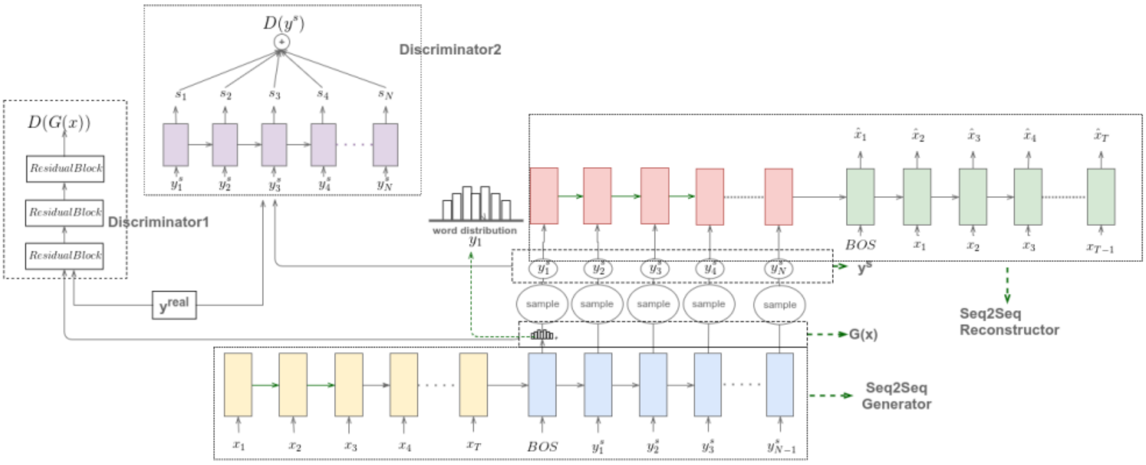
$$d_i = \sum_{j=i}^{N} \gamma^{j-i} r_j^D .$$

To maximize the expected discounted reward d_i , the loss of generator is

$$G'_{loss} = -E_{y_i^s \sim p_G(y_i^s | y_1^s, \ldots, y_{i-1}^s, x)}[d_i]$$

We use the likelihood ratio trick to approximate the gradient to minimize.

10. Architecture of the synthetic headline generation model



$D(y^s)$
Discriminator2
s_1 s_2 s_3 s_4 s_N
y_1^s y_2^s y_3^s y_4^s y_N^s
$D(G(x))$
ResidualBlock
ResidualBlock
ResidualBlock
Discriminator1
y^{real}
word distribution y_1
$\hat{x}_1$ $\hat{x}_2$ $\hat{x}_3$ $\hat{x}_4$ $\hat{x}_T$
BOS x_1 x_2 x_3 x_{T-1}
y_1^s y_2^s y_3^s y_4^s y_N^s
y^s
sample
sample
sample
sample
sample
G(x)
Seq2Seq Reconstructor
Seq2Seq Generator
x_1 x_2 x_3 x_4 x_T BOS y_1^s y_2^s y_3^s y_{N-1}^s

11. Inter-mutual Attentive Semantic Matching

The original headline and synthetic headlines are represented using vectors obtained after word embedding.Further the proposed model is performing three tasks:

1. Word to Word Inter Mutual Attention using Difference
2. Word to Word Inter Mutual Attention using Dot-Product
3. Word to Word Inter Mutual Attention using Concatenation

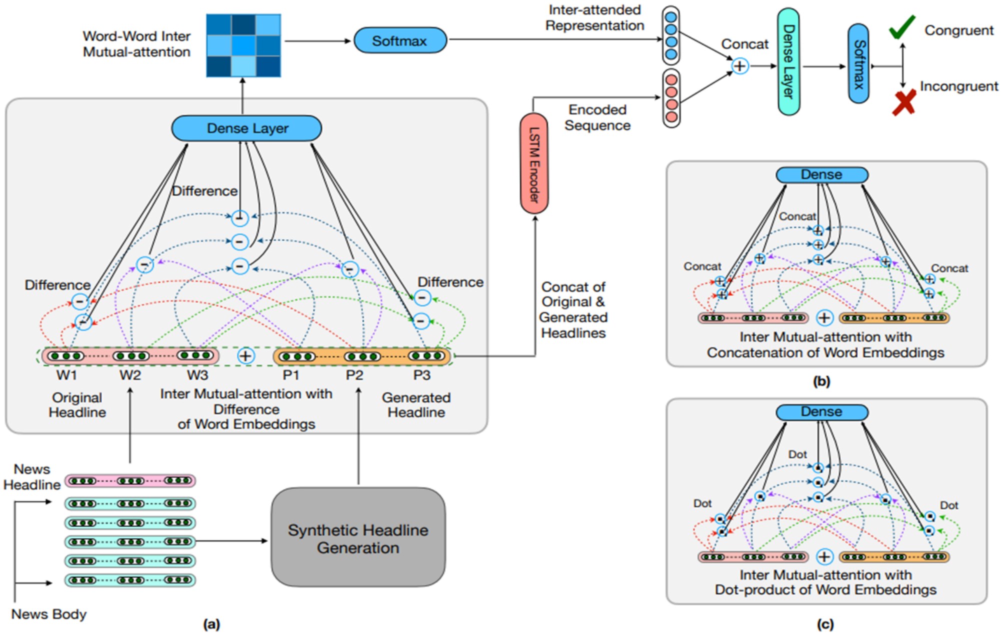
We then compute inter-mutual attended representations for original headlines and for synthetic headlines.

$$M_{A^o} = \sum_{i=1}^{l} f(w_{h^o i}) A^{o}{}_{i}$$

$$M_{A^s} = \sum_{i=1}^{p} f(W_{h^s i}) A^{s}{}_{i},$$

$$M_A = M_{A^o} + M_{A^s}$$

12. Architecture Of The Proposed Model


Word-Word Inter Mutual-attention
Softmax
Inter-attended Representation
Concat
Dense Layer
Softmax
Congruent
Incongruent
Encoded Sequence
LSTM Encoder
Dense Layer
Difference
Difference
Difference
W1
W2
W3
P1
P2
P3
Original Headline
Inter Mutual-attention with Difference of Word Embeddings
Generated Headline
Concat of Original & Generated Headlines
News Headline
News Body
Synthetic Headline Generation
(a)
Dense
Concat
Concat
Concat
Inter Mutual-attention with Concatenation of Word Embeddings
(b)
Dense
Dot
Dot
Dot
Inter Mutual-attention with Dot-product of Word Embeddings
(c)

13. Clubbed Model

$$F_{qr}^{dot} = [f(w_{h^o q}) \cdot f(w_{h^s r})]$$

$$F_{qr}^{con} = [f(w_{h^o q}) \,\|\, f(w_{h^s r})]$$

$$F_{qr}^{diff} = [f(w_{h^o q}) - f(w_{h^s r})]$$

$$C_{qr} = \theta_{dpc}([F_{qr}^{dot} \,\|\, F_{qr}^{con} \,\|\, F_{qr}^{diff}]) + b_{dpc},$$

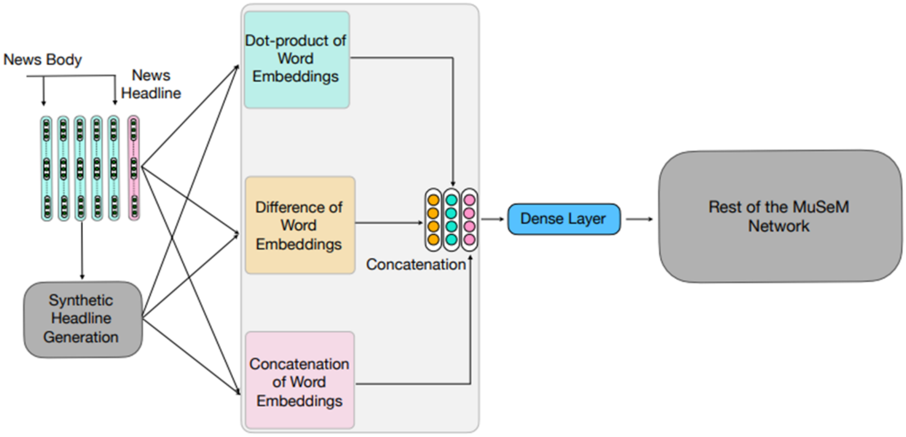

News Body
News Headline
Synthetic Headline Generation
Dot-product of Word Embeddings
Difference of Word Embeddings
Concatenation of Word Embeddings
Concatenation
Dense Layer
Rest of the MuSeM Network

Figure 3: Clubbed Model Architecture

14. Dataset

Statistics	NELA
Total	71420
Non-Congruent	35710
Congruent	35710

15. Results

Accuracy	0.737374
F1 Score	0.666667

The above mentioned results were obtained on running the model over NELA dataset.

16. Conclusion

We have implemented inter-mutual attention based semantic matching (MuSeM) to detect incongruence in news headlines, which combines three operations on word embedding pairs to compute inter-mutual attention scores.We notice that the performance of inter-mutual attention based semantic matching greatly depends on the accuracy of synthetic headline generation step.

17. References

1. https://arxiv.org/abs/2010.03617
2. https://arxiv.org/abs/1810.02851