Detecting Incongruent News Articles Using Multi-head Attention Dual Summarization

Sujit Kumar, Gaurav Kumar, Sanasam Ranbir Singh
Department of Computer Science and Engineering
Indian Institute of Technology, Guwahati, Assam, India
{sujitkumar,ranbir}@iitg.ac.in, gauravchaudhary216@gmail.com

Abstract

With the increasing use of influencing incongruent news headlines for spreading fake news, detecting incongruent news articles has become an important research challenge. Most of the earlier studies on incongruity detection focus on estimating the similarity between the headline and the encoding of the body or its summary. However, most of these methods fail to handle incongruent news articles created with embedded noise. Motivated by the above issue, this paper proposes a *Multi-head Attention Dual Summary* ($MADS$) based method which generates two types of summaries that capture the congruent and incongruent parts in the body separately. From various experimental setups over three publicly available datasets, it is evident that the proposed model outperforms the state-of-the-art baseline counterparts.

1 Introduction

News headlines greatly influence opinion of the readers (Tannenbaum, 1953) and play a significant role in making a new viral on any social media (Rieis et al., 2015) (Gabielkov et al., 2016) (Wei and Wan, 2017). A deceitful and incongruent news article can negatively affect readers, such as false beliefs and wrong opinions [1][2] (Ecker et al., 2014) (Ecker et al., 2022) (Tsfati et al., 2020). If a news headline misrepresents the content of its body then such headline and body pair is called incongruent news article (Chesney et al., 2017) (Wei and Wan, 2017). In recent times, usage of deceptive and incongruent news headlines as an effective means to spread disinformation over digital platforms is evident (Chesney et al., 2017) (Effron and Raj, 2020) [3][4]. Consequently, detecting deceitful and incongruent news articles (Chesney et al., 2017) (Ecker et al., 2014) (Horner et al., 2021) (Bago et al., 2020) (Guess et al., 2020) is becoming an important research problem to counter the spread of misinformation over digital media.

An incongruent news article may be constituted in various forms (i) the headline makes unrelated or opposite claims to its body, (ii) both headline and body refer to a common topic or event, but the contents are not related, (iii) both headline and body report a genuine event/incident, but the dates or name entities are manipulated, (iv) methods are Earlier studies on incongruent news detection mainly focuses on estimating dissimilarity between headline and body using methods such as bag-of-words based features (Pomerleau and Rao, 2017), (Hanselowski et al., 2017), (Riedel et al., 2017), sequential encoding of headline and body (Hanselowski et al., 2018), (Borges et al., 2019), and hierarchical encoding of the news article (Karimi and Tang, 2019), (Conforti et al., 2018), (Yoon et al., 2019). As reported in (Mishra et al., 2020), the above similarity-based methods generally fail to detect incongruent news for the news article body with larger paragraphs and sentences. To address these problems, recent studies (Sepúlveda-Torres et al., 2021), (Mishra et al., 2020), (Kim and Ko, 2021a) propose summarization-based approaches. As the summarization in these studies are biased towards the dominant content of the body, such summarization may fail to capture the embedding noise present in partially incongruent news articles. Motivated by this, this paper proposes a *Multi-head Attention Dual Summarization* $MADS$ based summarization method which is capable of handling partially incongruent news by summarizing both the congruent and incongruent part of the article body. The proposed method divides the body of the news article into two sets - *positive: highly congruent sentences with headline* and *negative: highly incongruent sentences with headline*. Further, for each set, different forms of representation are cap-

[1] Impact of misleading headline in health
[2] Misleading headlines effect on economy news
[3] Examples of misleading headline fake news
[4] Misleading headline fake news over WHO

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 967–977
November 20–23, 2022. ©2022 Association for Computational Linguistics

tured using multi-head attention and convolution. From various experiments over three publicly available benchmark datasets, it is observed that the proposed method outperforms the existing state-of-the-art baseline counterparts, including the dataset with partially incongruent news article.

2 Related Work

Though both the clickbait and incongruent news article detection relate to news headline, as discussed in (Park et al., 2020), (Chesney et al., 2017), clickbait headline can be detected based on the headline only, whereas incongruent news article is defined by the relation between the headline and the news article body (Park et al., 2020). Clickbait attempts to attract the reader's attention, but incongruent news articles do not force readers to click some link and follow up (Chesney et al., 2017). Our paper focuses on incongruent detection. Studies on incongruent news article detection can be broadly categorized into similarity-based and summarization-based approaches. Initial studies (Pomerleau and Rao, 2017), (Hanselowski et al., 2017), (Riedel et al., 2017) (Hanselowski et al., 2018), (Borges et al., 2019) (Bhatt et al., 2018)used bag-of-word based features and sequential encoding to discover similarity between headline and body to detect incongruity. Further studies under similarity-based approaches exploit attention between headline and body (Conforti et al., 2018) (Mohtarami et al., 2018) (Saikh et al., 2019) (Jang et al., 2022) for incongruent news article detection. Studies (Karimi and Tang, 2019) (Yoon et al., 2019), (Yoon et al., 2021) utilize hierarchical structure of news article to highlight important sentences in body with respect to claim of headline. However, the similarity-based approach performs average when the news article body is significantly large (high number of words and sentences) compared to the headline's length (Mishra et al., 2020), (Sepúlveda-Torres et al., 2021). Also, similarity-based methods fail to detect partially incongruent news articles. To overcome the limitations of the similarity-based approach, studies (Mishra et al., 2020), (Sepúlveda-Torres et al., 2021) make use of the summarization technique to summarize news articles body to pieces of text. Subsequently, text matching methods are applied between the summary of the news article body and the headline. Studies (Kim and Ko, 2021a) (Kim and Ko, 2021b) exploit graph summarization to detect fake news articles. Study (Mishra and Zhang, 2021) make use of Part of Speech tag patterns(POS) based attention to take cognizance of numerical value of headlines and body for incongruent news article detection. Considering the importance of bidirectional context in documents, study (Kumar et al., 2022) propose RoBERT-based models for fake news detections. A recent study (Jang et al., 2022) utilizes news subtitle, image caption, headline and body along with attention between headline and body to detect incongruent headline.

As the summarization in these studies are biased towards the dominant content of the body, such summarization may fail to capture the embedding noise present in partially incongruent news articles. Hence, we need an incongruent news article detection-specific summarization technique, which should focus more on the incongruent part of the news article while generating a summary of news article body. Considering such limitations of summarization-based approach for incongruent news detection, this paper proposes a Multi-head Attention Dual Summarization model $MADS$ which divide the body into two sets : positive set and negative set. If the similarity score of a sentence with the headline is high, then it is placed in a positive set and otherwise placed in a negative set. Then a summary of both sets is obtained separately and matched with the headline for incongruent news article detection.

3 Proposed Models

Given a news article $\mathcal{I} = \left(\mathcal{H}, \mathcal{B}\right)$ with a pair of its headlines $\mathcal{H}$ and its body $\mathcal{B}$, $MADS$ divides the sentences in the body $\mathcal{B}$ into positive $\mathcal{P}$ and negative $\mathcal{N}$ sets based on the matching scores between the sentence $\mathcal{S}_i$ and the headline $\mathcal{H}$. The main motivation behind splitting body sentences into positive $\mathcal{P}$ and negative $\mathcal{N}$ sets is that if a news article is partially incongruent, then sentences congruent with the headline will be in positive set $\mathcal{P}$ and sentences incongruent with a headline will be in negative set $\mathcal{N}$. Similarly, in the case of a full congruent news article, most of the sentences of the body should be in $\mathcal{P}$ set, and only few sentences will be in $\mathcal{N}$ set. However, if a news article is fully incongruent, then all the sentences in the body should be incongruent with the headline; hence it should be in $\mathcal{N}$ except one or few sentences in $\mathcal{P}$. Next, summary of $\mathcal{P}$ and $\mathcal{N}$ are obtained separately to match with

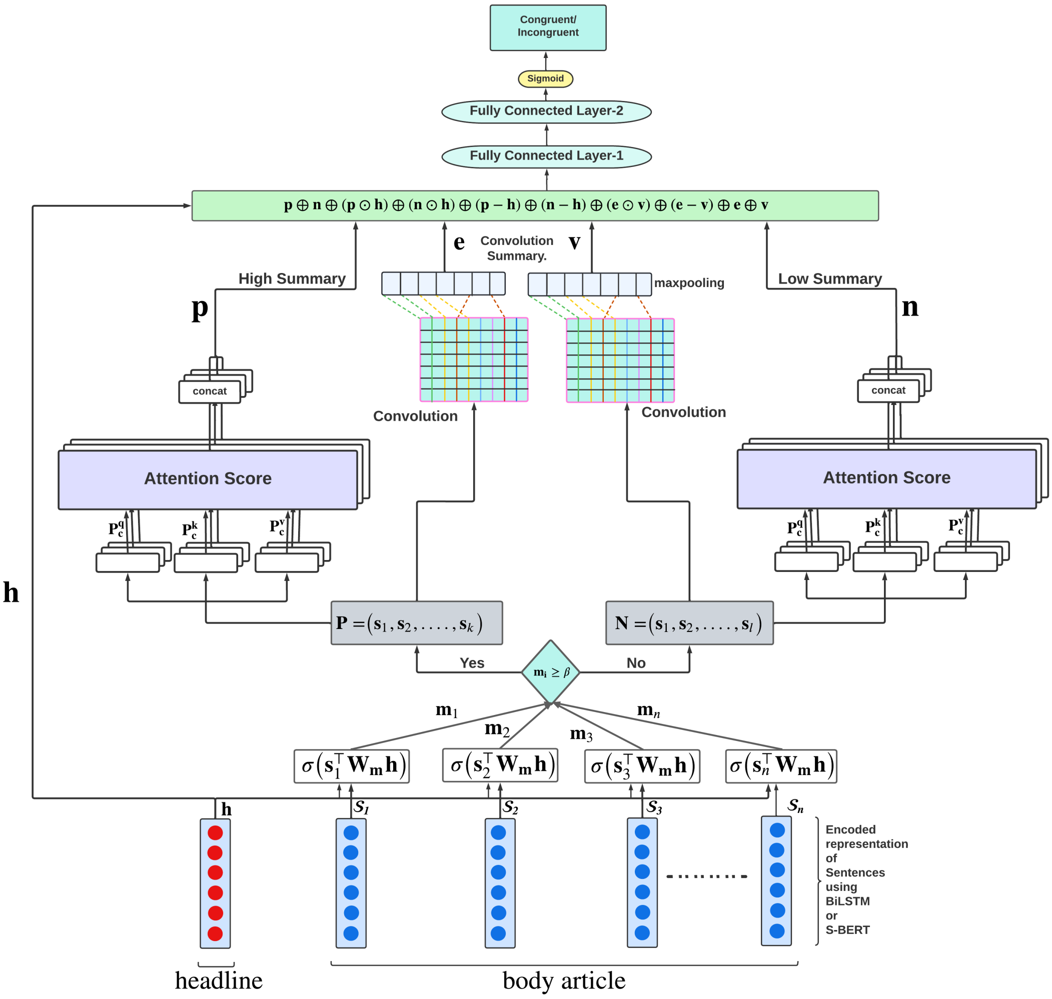
Congruent/ Incongruent
Sigmoid
Fully Connected Layer-2
Fully Connected Layer-1
$\mathbf{p} \oplus \mathbf{n} \oplus (\mathbf{p} \odot \mathbf{h}) \oplus (\mathbf{n} \odot \mathbf{h}) \oplus (\mathbf{p} - \mathbf{h}) \oplus (\mathbf{n} - \mathbf{h}) \oplus (\mathbf{e} \odot \mathbf{v}) \oplus (\mathbf{e} - \mathbf{v}) \oplus \mathbf{e} \oplus \mathbf{v}$
High Summary
e
Convolution Summary.
v
maxpooling
Low Summary
p
n
concat
concat
Convolution
Convolution
Attention Score
Attention Score
$\mathbf{P}_c^q$ $\mathbf{P}_c^k$ $\mathbf{P}_c^v$
$\mathbf{P}_c^q$ $\mathbf{P}_c^k$ $\mathbf{P}_c^v$
h
$\mathbf{P} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_k)$
$\mathbf{N} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_l)$
Yes
$\mathbf{m}_i \geq \beta$
No
$\mathbf{m}_1$
$\mathbf{m}_2$
$\mathbf{m}_3$
$\mathbf{m}_n$
$\sigma(\mathbf{s}_1^\top \mathbf{W}_\mathbf{m} \mathbf{h})$
$\sigma(\mathbf{s}_2^\top \mathbf{W}_\mathbf{m} \mathbf{h})$
$\sigma(\mathbf{s}_3^\top \mathbf{W}_\mathbf{m} \mathbf{h})$
$\sigma(\mathbf{s}_n^\top \mathbf{W}_\mathbf{m} \mathbf{h})$
h
$\mathcal{S}_1$
$\mathcal{S}_2$
$\mathcal{S}_3$
$\mathcal{S}_n$
Encoded representation of Sentences using BiLSTM or S-BERT
headline
body article

Figure 1: The proposed model $MADS$ is represented in the diagram. First, sentence encoding are obtained using BiLSTM or S-BERT. Then, a similarity score m_i between $\mathbf{h}$ and $\mathbf{s}_i$ is estimated. If $m_i \geq \boldsymbol{\beta}$ is true, the sentence is placed in the positive set otherwise, it is placed in the negative set. Then we generate summary of these positive and negative set using multi-head attention and convolution. Thereafter, text matching features between headline and representative summary generated from multi-head attention and convolution is obtained and passed to the two fully connected layers for the classification.

headline for incongruent news article detection.

3.1 Similarity Between Headline and Body:

This study uses bidirectional LSTM (BiLSTM) to obtain encoded representation $\mathbf{h}$ and $\mathbf{s}_i$ of headline $\mathcal{H}$ and sentence $\mathcal{S}_i$, respectively. However, considering the effectiveness of sentence embeddings generated by sentence-BERT (S-BERT) (Reimers and Gurevych, 2019) in different NLP tasks[5], we have also used S-BERT to encode headline and sentences, in this study. Like in (Tay et al., 2018) (Luong et al., 2015), the similarity score m_i between $\mathbf{h}$ and $\mathbf{s}_i$ is estimated using the following expression 1

$$m_i = \sigma\left(\mathbf{s}_i^\top \mathbf{W}_\mathbf{m} \mathbf{h}\right) \tag{1}$$

where $\mathbf{W}_\mathbf{m}$ is a learnable parameter matrix, σ is the sigmoid function and $\top$ is a transpose operation over a vector. If $m_i \geq \boldsymbol{\beta}$, then sentence $\mathbf{s}_i$ is added to set $\mathcal{P}$, otherwise it is added to set $\mathcal{N}$.

3.2 Summarization

Given two sets of sentences, $\mathcal{P}$ and $\mathcal{N}$, we extract two different types of summaries - *multi-head attention-based summary* and *convolution summary* for each set separately.

[5] Why S-BERT

3.2.1 Summary using Multi-head Attention

The characteristics of dual summary over positive $\mathcal{P}$ and negative $\mathcal{N}$ sets are defined as follows: *(i)* a sentence which is highly similar to other sentences in the set $\mathcal{P}$ should be given high priority while generating a summary of a positive set $\mathcal{P}$. *(ii)* A sentence which is not similar or least similar to other sentences in the set $\mathcal{N}$ should be given high importance while generating a summary of $\mathcal{N}$. The main motivation behind such a dual summary is that if a summary generated by a highly influenced (sentence with high similarity with all other sentences in the set) sentence from a positive set and a summary generated by the least influenced (a sentence which is either not similar or least similar with other sentences in the set) sentence from $\mathcal{N}$ are congruent with the headline, then the news article is congruent, otherwise incongruent. To capture representation of sentences from different aspects, we apply multi-head attention (Vaswani et al., 2017). As shown in Figure 1, given a sequence of sentences $(\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_k)$, we define a matrix $\mathbf{P}$ (each row representing a sentence encoding) to obtain the query $\mathbf{P}^q$, key $\mathbf{P}^k$ and value $\mathbf{P}^v$ matrices using the following expression.

$$\mathbf{P}^q_c, \mathbf{P}^k_c, \mathbf{P}^v_c = \mathbf{P} \cdot \mathbf{W}^q_c, \mathbf{P} \cdot \mathbf{W}^k_c, \mathbf{P} \cdot \mathbf{W}^v_c \quad (2)$$

where $\mathbf{W}^q_c$, $\mathbf{W}^k_c$ and $\mathbf{W}^v_c$ are learnable parameter matrices of query, key and value projections respectively, for c^{th} attention head of multi-head self attention and $\cdot$ is the dot product between matrix. Subsequently, attention weigh $\mathbf{A}_c$ is defined as follows:

$$\mathbf{M} = \left(\frac{\mathbf{P}^q_c\,(\mathbf{P}^k_c)^\top}{\sqrt{\mathbf{z}}}\right) \quad (3)$$

$$\mathbf{A}_{c,i,j} = \left(\frac{\mathbf{exp}(\mathbf{M_{ij}})}{\sum_{k,l}\mathbf{exp}(\mathbf{M_{k,l}})}\right) \quad (4)$$

Here M is matching matrix and $\mathbf{A}_c$ is attention weight matrix of c^{th} attention head. $\mathbf{A_c}[\mathbf{i},\mathbf{j}]$ entry represents the similarity probability between i^{th} and j^{th} sentence of set $\mathcal{P}$. $\mathbf{z}$ is the dimension of $\mathbf{P}^q_c$. Next, weighted summation is applied over encoding of sentences $\mathbf{s}_i$ based on similarity with other sentences in the set.

$$\mathbf{u}_{c,i} = \left(\sum_{j=1,i\neq j}^{k} \mathbf{A}_{c,ij}\mathbf{P}^v_{c,i}\right) \quad (5)$$

Where $\mathbf{u}_{c,i}$ is the sentence representation obtained after weighted summation between i^{th} sentence of $\mathbf{P}^v_c$ and attention weight $\mathbf{A}_{c,ij}$ between i^{th} sentence with all other sentences j in $\mathbf{P}^v_c$ of attention head c. Similarly, by following equation 5, representation of other sentences in a respective set are also obtained to form a sentence representation matrix $\mathbf{U}_c = \{\mathbf{u}_{c,1}, \mathbf{u}_{c,2}, ..., \mathbf{u}_{c,k}\}$ of attention head c. Now we concatenate the sentence representation obtained by different attention head and pass it to dense layer to obtained final sentence representation $\mathbf{U}$.

$$\mathbf{U} = \Big(\mathbf{U}_1 \oplus \mathbf{U}_2 \oplus ..\mathbf{U}_c \oplus . \oplus \mathbf{U}_l\Big)\mathbf{W}_u \quad (6)$$

Where $\mathbf{W}_u$ is the trainable parameter matrix and $\mathbf{U}_c$ is c^{th} attention head. $\mathbf{U}$ is sentence representation matrix obtained by concatenating representation of i^{th} sentence obtained by l attention head. Now we concatenate representations of sentences $\mathbf{u}_i$ in the sentence representation matrix $\mathbf{U}$ and pass to dense layer to obtain a summary $\mathbf{p}$ of positive set $\mathcal{P}$.

$$\mathbf{p} = \Big(\mathbf{u}_1 \oplus \mathbf{u}_2 \oplus .. \oplus \mathbf{u}_i \oplus . \oplus \mathbf{u}_k\Big)\mathbf{W}_m \quad (7)$$

Where $\mathbf{u}_i$ is a row vector of the matrix U and $\mathbf{W}_m$ is the learnable parameter matrix. Similarly, to extract a summary $\mathbf{n}$ of a negative set, $\mathcal{N}$ equation 4 is replaced by equation 8. The reason behind this is that the sentence with the least similarity score with other sentences in the set $\mathcal{N}$ should be given high importance while generating a summary $\mathbf{n}$ of set $\mathcal{N}$.

$$\mathbf{A}_{c,i,j} = \left(\frac{\mathbf{exp}(\mathbf{1} - \mathbf{M_{ij}})}{\sum_{k,l}\mathbf{exp}(\mathbf{1} - \mathbf{M_{k,l}})}\right) \quad (8)$$

3.2.2 Local Patterns Summary

We also extract a summary by extracting meaningful n-grams substructure and local patterns within sentence encoding matrix $\mathbf{P}$ and $\mathbf{N}$ of positive set $\mathcal{P}$ and negative $\mathcal{N}$ sets respectively. To extract summary $\mathbf{e}$ and $\mathbf{v}$ based on the local structure and meaningful n-grams substructure, we employ convolution (Kim, 2014) over positive $\mathcal{P}$ and negative $\mathcal{N}$ sets. Our convolution settings over sentence encoding matrix $\mathbf{P}$ and $\mathbf{N}$ of positive $\mathcal{P}$ and negative $\mathcal{N}$ sets are similar to convolution setting discussed in study (Kim, 2014)[6]. We concatenate the summary obtained by unigrams, bigrams, trigrams upto 7-grams convolution operations to generate summary $\mathbf{e}$ and $\mathbf{v}$ of positive $\mathcal{P}$ and negative $\mathcal{N}$ sets respectively.

[6] Convolutional Neural Networks Implementation GitHub Link

Subsequently, we further estimate feature vectors to measure similarity and contradiction between headline encoding $\mathbf{h}$ and summary obtained using multi-head attention $\mathbf{p}$, $\mathbf{n}$. The main objective behind estimating similarity and contradiction between headline and summary of the positive and negative set is that if a news article is fully congruent, then the similarity between the headline and summary of positive and negative sets should be high. Similarly, in the case of fully incongruent news article, the similarity of headline encoding $\mathbf{h}$ with both summaries $\mathbf{p}$ and $\mathbf{n}$ should be low. Intuitively, in the case of a partially incongruent news article, the similarity between headline encoding $\mathbf{h}$ and summary $\mathbf{p}$ of the positive set may be high. Still, the similarity between headline encoding $\mathbf{h}$ and summary $\mathbf{n}$ of negative set should be low. With the above-mentioned objectives, we estimated similarity and contradiction between headline and summary of positive and negative set as follows:

$$\mathbf{a}^{+} = \mathbf{p} \odot \mathbf{h} \tag{9}$$

$$\mathbf{a}^{-} = \mathbf{n} \odot \mathbf{h} \tag{10}$$

$$\mathbf{b}^{+} = \mathbf{p} - \mathbf{h} \tag{11}$$

$$\mathbf{b}^{-} = \mathbf{n} - \mathbf{h} \tag{12}$$

$$\acute{\mathbf{f}} = \left(\mathbf{a}^{+} \oplus \mathbf{a}^{-} \oplus \mathbf{b}^{+} \oplus \mathbf{b}^{-} \oplus \mathbf{p} \oplus \mathbf{n}\right) \tag{13}$$

Where $\odot$ denotes element-wise multiplication and $\oplus$ denotes concatenation of vectors. $\mathbf{a}^{+}$ and $\mathbf{b}^{+}$ is angle and difference (similarity measure features) between summary of positive set and headline. Similarly, $\mathbf{a}^{-}$ and $\mathbf{b}^{-}$ are similarity feature between headline and summary of negative set. Next, we also estimate the similarity between $\mathbf{e}$ and $\mathbf{v}$ convolution summary of positive set $\mathcal{P}$ and negative set, $\mathcal{N}$ respectively. The key motivations behind estimating similarity between $\mathbf{e}$ and $\mathbf{v}$ is that if a news article is congruent, then similarity between the summary of positive set $\mathcal{P}$ and negative set $\mathcal{N}$ should be high because sentences in the body of a congruent news article are related to each other and similar in topics. Whereas in case of partially incongruent or fully incongruent article, there must be some sentences in body content which does not correlate with headline and other sentences of body. Hence, in case of incongruent news article, dissimilarity between summary of positive set $\mathcal{P}$ and negative set $\mathcal{N}$ should be high. With such motivation, we estimate similarity between $\mathbf{e}$ and $\mathbf{v}$ convolution summary of positive set $\mathcal{P}$ and negative set $\mathcal{N}$ as follows:

$$\mathbf{c}^{+} = \mathbf{e} \odot \mathbf{v} \tag{14}$$

$$\mathbf{c}^{-} = \mathbf{e} - \mathbf{v} \tag{15}$$

$$\mathbf{f} = \left(\acute{\mathbf{f}} \oplus \mathbf{c}^{+} \oplus \mathbf{c}^{-} \oplus \mathbf{e} \oplus \mathbf{v}\right) \tag{16}$$

Finally, the feature vector $\mathbf{f}$ is passed to a two-layer fully connected neural network followed by softmax for incongruent news article classification.

Table 1: Characteristics of Experimental Datasets

	Dataset	Cong.	Incong.	Total	#Head	#Body	#Para	#Sen
ISOT	Train	17083	18232	35315	9.438	244.325	3.799	16.955
	Test	1726	1815	5313	9.377	236.379	3.729	16.606
	Dev	2607	2706	3541	9.388	241.136	3.733	16.607
FNC	Train	40321	15161	55482	11.133	361.326	10.782	19.113
	Test	11039	4038	15077	8.503	365.027	10.950	19.331
	Dev	3533	1292	4825	11.174	363.417	10.916	19.203
NELA-17	Train	35710	35710	71420	10.558	551.923	13.494	26.649
	Test	3151	3151	6302	10.529	566.921	13.851	27.526
	Dev	3151	3151	6302	10.547	541.188	13.49	26.256

4 Experimental Results and Discussions

4.1 Dataset

This study considers three publicly available datasets of different natures, namely the ISOT fake news dataset [7] [8] (Ahmed et al., 2018) (Ahmed et al., 2017), Fake News Challenge (FNC) dataset[9] (Pomerleau and Rao, 2017), and NELA-17 (News Landscape) dataset (Horne et al., 2018), (Yoon et al., 2019). The FNC dataset has four classes, namely: agree, disagree, discuss, and unrelated. Samples from agree, disagree and discuss classes are merged and named as a congruent *Cong.* class, whereas the samples in unrelated class are considered as incongruent *Incong.* class. An important characteristic of the FNC dataset is that the samples in the unrelated (fake) are generated by taking headlines and bodies from two different news articles under different topics (Hanselowski et al., 2018). We therefore refer the samples under unrelated class as fully incongruent news articles. We curate NELA dataset by following the procedure[10] reported in study (Yoon et al., 2019) over news article corpus[11] released by study (Horne et al., 2018). As reported in study (Yoon et al., 2019) news articles published by authentic media house are considered as congruent *Cong.*, whereas

[7] ISOT: Information Security and Object Technology (ISOT)

[8] ISOT Fake News Dataset Repository Source

[9] Fake News Challenge (FNC)

[10] NELA Dataset Generator Procedure and Code

[11] NELA-17 Dataset News Article Corpus

incongruent *Incong.* news articles are generated, inserting a paragraph from a randomly selected news article into *Cong.* news article. Since a paragraph is inserted into a *Cong.* news article to generate *Incong.* samples, it is obvious that all other paragraph except which is inserted will be congruent with the headline. Hence, *Incong.* samples in NELA dataset are partially incongruent. ISOT dataset (Ahmed et al., 2018) (Ahmed et al., 2017) is curated by considering news articles published by authenticated source as class samples, whereas news articles published by unverified or unauthenticated source are considered as *False* class samples. NELA and ISOT datasets are balanced datasets, but FNC dataset is an imbalanced dataset.

4.2 Experimental Setups

To compare the performance of the proposed model, we consider several existing state-of-the-art models from the literature as baselines. These baselines models can be grouped into two categories: *(i)* Similarity-based methods, *(ii)* Summarization-based methods.

Similarity-based methods: This paper considers bag-of-words features-based methods `FNC` (Fake News Challenge) (Pomerleau and Rao, 2017), `UCLMR` (UCL Machine Reading) (Riedel et al., 2017). We consider encoding-based methods `StackLSTM` (Hanselowski et al., 2018), `HDSF` (Hierarchical Discourse level Structure Learning) (Karimi and Tang, 2019), `AHDE` (Attentive Hierarchical Dual Encoder) (Yoon et al., 2019) `GHDE` (Graph-based Hierarchical Dual Encoder) (Yoon et al., 2021) as baselines. The default settings and codes available at their respective GitHub code repository `FNC`[12] `UCLMR`[13] `stackLSTM`[14] `HDSF`[15] `AHDE`[16] `GHDE`[17] have been used to reproduce the results. As `GHDE` models needs paragraph level annotations, it has been tested only with NELA dataset, where the inserted paragraphs are annotated as incongruent. **Summarization-based methods:** This paper considers a recent study `FEDS` (Fake news Detection using Summarization) (Kim and Ko, 2021b) (Kim and Ko, 2021a) as summarization-based baseline.

Apart from the similarity and summarization-based baseline discussed above, we consider other four different baselines.

`BiLSTM`: This model finds entailment and similarity between headline and body content to decide congruence between headline and body. First, the headline and body are encoded using BiLSTM (Hochreiter and Schmidhuber, 1997). Next, the angle and difference between encoded headline and body are concatenated with the encoded representation of headline and body to form an entailment feature. Finally, the entailment feature is passed to a fully connected neural network, followed by Softmax for incongruent news article classifications.

`BERT`: This baseline model follows a similar approach to BiLSTM, except it use pretrained BERT[18] (Devlin et al., 2019) to encode headline and body.

`RoBERT`: (Recurrence over BERT) (Pappagari et al., 2019) This is hierarchical transformer model which first split news article into several sentences. Then, encoding of each sentence is obtained using pretrained BERT (Devlin et al., 2019). Subsequently, RoBERT model, applies an LSTM over the encoding of sentences to obtain encoding of the body. Finally, the encoding of the body is passed to a fully connected neural network for incongruent news classifications. LSTM is applied over the encoding of sentences with intuitions that a news article is a sequence of sentences and each sentence is related to the next and previous sentence.

`MAS`: (Multi-head Attention Summarization) It is similar to the proposed model $MADS$, but does not split the news article body into two sets for summarizations. Instead, it applies multi-head attention and convolution summarization over full-body contents. All other settings are similar to the proposed model $MADS$.

We use Google's word2vec (Mikolov et al., 2013) pre-trained embedding for word level embedding. The F-measure (F), classwise F-measure, Accuracy (Acc) have been used as evaluation metrics. The details of experimental hyperparameters are present in A. Our code repository is publicly available[19] `https://github.com/thesujitkumar/Multi_`

[12] FNC-1 baseline by organizer code

[13] UCLMR implementation code

[14] stackLSTM based model code repository

[15] HDSF code repository

[16] Attentive Hierarchical Dual Encoder(AHDE) code

[17] GHDE model code repository

[18] Huggingface pretrained BERT

[19] `https://github.com/thesujitkumar/Multi_Head_Attention_Dual_Summarization.git`

Table 2: Comparison of the performances of different models over three benchmark datasets. Here, (Acc) and (F) indicate accuracy and F-measure, respectively. Similarly, (Cong.) and (Incong.) indicate F-measure of congruent and incongruent class, respectively.

			NELA-17				ISOT				FNC			
		Models	Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.	Acc	F	Cong.	Incong.
Baseline	**Feat.**	**FNC** (Pomerleau and Rao, 2017)	0.586	0.586	0.564	0.608	0.844	0.844	0.847	0.842	0.586	0.496	0.282	0.709
		UCLMR (Riedel et al., 2017)	0.589	0.588	0.608	0.569	0.997	0.997	0.997	0.997	0.964	0.955	0.934	0.975
		StackLSTM (Hanselowski et al., 2018)	0.597	0.591	0.541	0.641	0.992	0.992	0.992	0.992	**0.971**	**0.963**	**0.946**	**0.982**
	Encoding	**AHDE** (Yoon et al., 2019)	0.606	0.606	0.614	0.598	0.913	0.913	0.909	0.909	0.691	0.454	0.094	0.814
		HDSF (Karimi and Tang, 2019)	0.517	0.494	0.602	0.386	0.720	0.712	0.665	0.759	0.758	0.666	0.492	0.841
		GHDE (Yoon et al., 2021)	0.55	0.331	0.331	0.332	-	-	-	-	-	-	-	-
		FEDS (Kim and Ko, 2021b) (Kim and Ko, 2021a)	0.533	0.532	0.550	0.515	**0.998**	**0.998**	**0.998**	**0.998**	0.878	0.837	0.755	0.918
		BiLSTM	0.555	0.55	0.563	0.547	0.99	0.99	0.99	0.99	0.616	0.504	0.269	0.74
		BERT	0.572	0.563	**0.624**	0.503	0.894	0.894	0.894	0.891	0.722	0.419	0.21	0.838
		RoBERT	**0.615**	**0.613**	0.54	**0.642**	0.996	0.996	0.996	0.996	0.664	0.583	0.4	0.767
		MAS	0.543	0.528	0.445	0.611	0.997	0.997	0.997	0.997	**0.958**	**0.947**	**0.923**	**0.971**
Proposed	**Encoding**	**MADS(BiLSTM, β = 0.5 , H = 8)**	0.581	0.575	0.527	0.623	**0.999**	**0.999**	**0.999**	**0.999**	**0.971**	**0.963**	**0.947**	**0.98**
		MADS(BiLSTM, β = 0.5 , H = 2)	0.624	0.623	0.637	0.609	0.998	0.998	0.998	0.998	0.966	0.958	0.939	0.977
		MADS(BiLSTM, β = 0.5 , H = 1)	**0.641**	**0.640**	**0.652**	**0.629**	0.998	0.998	0.998	0.998	0.969	0.960	0.942	0.978
		MADS(S-BERT, β = 0.5 , H = 1)	0.63	0.628	0.603	0.654	0.984	0.984	0.984	0.984	**0.971**	**0.963**	**0.947**	**0.98**
		MADS(S-BERT, β = 0.5 , H = 2)	0.625	0.62	0.579	0.662	0.972	0.972	0.972	0.972	0.968	0.959	0.94	0.978
		MADS(S-BERT, β = 0.5 , H = 8)	0.568	0.562	0.514	0.593	0.978	0.977	0.977	0.978	0.962	0.952	0.93	0.974

`Head_Attention_Dual_Summarization.git` to reproduce the results of our proposed model setup.

4.3 Results and discussion

Table 2 presents the comparison between the performance of baselines and proposed models over three benchmark datasets. As discussed in section 4.1, due to different characteristics possessed by the three datasets, proposed and baseline models respond differently to them. First, we study the performance of baseline models, which are divided into *explicit* and *neural encoding*, depending on whether a model uses explicit features or neural models to encode news headlines and body. Feature-based models outperform neural encoding-based models over FNC dataset, while for NELA and ISOT datasets, their performance is comparable. Summarization-based methods MAS and $FEDS$ outperform neural encoding models over FNC and ISOT datasets. This indicates that matching between summary of news article body and headline is more effective than matching between headline and global encoding of body. However, $RoBERT$ outperforms MAS and $FEDS$ over the NELA dataset. This indicates that summarization-based methods are effective only in case of incongruent news detection, but performs poorly for partially incongruent news detections. Our proposed model **MADS** attempts to overcome the limitation of summarization-based methods for partially incongruent news detection by generating a multi-head attention dual summary.

Table 2 presents different setups of **MADS** differing in three parameters: *(i)* encoding headline and body sentences using BiLSTM (Hochreiter and Schmidhuber, 1997) or sentence BERT (S-BERT) (Reimers and Gurevych, 2019), *(ii)* H denotes number of head in multi-head attention summarization. These different setups are named as $MADS(BiLSTM, \beta, H)$ and $MADS(S-BERT, \beta, H)$ with different value of H and β in the Table 2. We consider three different values of H 1, 2 and 8. From table 2 it is apparent that $MADS(BiLSTM, \beta = 0.5, H = 8)$ and $StackLSTM$ jointly outperforms baseline models and other setup of proposed model over FNC dataset, however $MADS(BiLSTM, \beta = 0.5, H = 8)$ outperforms over ISOT dataset. From the performance of $MADS(BiLSTM, \beta = 0.5, H = 8)$ and $MADS(S-BERT, \beta = 0.5, H = 1)$ over FNC dataset, it can be claim that the value of H depend on sentence encoding methods. Similarly, $MADS(BiLSTM, \beta = 0.5, H = 1)$ outperforms baseline and other setup of proposed model over NELA dataset. From such observations, it establishes the superiority of our dual summary-based proposed model $MADS$ over baseline models for partially incongruent news article detection. To further validate this, we compare $MADS$ with summarization-based baseline models $FEDS$ and MAS. From table 2 it can be observed that $MADS$ outperform $FEDS$ (Kim and Ko, 2021a) (Kim and Ko, 2021b) and MAS over NELA, ISOT and FNC datasets. $MADS(BiLSTM, \beta = 0.5, H =$

1) outperform $FEDS$ and MAS by 20.26%, 18.047% over NELA dataset respectively. Similarly $MADS(BiLSTM, \beta = 0.5, H = 8)$ and $MADS(S-BERT, \beta = 0.5, H = 1)$ jointly outperform $FEDS$ and MAS by 10.59% and 1.38% over FNC dataset. These observations clearly establish the effectiveness of dual summarization over summarization-based incongruent news article detection. Thereafter, we compare summarization-based baselines $FESD$ and MAS, where MAS outperforms $FEDS$. This indicates that our proposed summarization method is more effective than the graph summarization approach of $FEDS$ (Kim and Ko, 2021a) (Kim and Ko, 2021b) for incongruent news article detection.

4.4 Dual Summary Versus Summary of Negative Set

Table 3: Comparison of the performances between Multi-head Attention Dual summarization $MADS$ and Multi-headed Attention and convolution-based Negative set Summarization $MANS$. Results are obtained using attention head $H = 1$ for NELA dataset and $H = 8$ for FNC and ISOT datasets.

	NELA		FNC		ISOT	
Model	Acc	F	Acc	F	Acc	F
MADS(BiLSTM , β = 0.5)	**0.641**	**0.64**	**0.97**	**0.963**	**0.999**	**0.999**
MANS(BiLSTM , β = 0.5)	0.619	0.618	0.927	0.907	0.997	0.997

$MADS$ estimates similarity between the headline and a summary of positive and negative set. Considering the essential characteristics of the negative set as discussed in section 3, It is intuitive to ignore the positive set summary and match the headline with the summary of the only negative set for incongruent news article detection. Table 3 present performance comparison between $MADS(BiLSTM, \beta = 0.5)$ and $MANS(BiLSTM, \beta = 0.5)$. $MANS$ (Multi-headed Attention and convolution-based Negative set Summarization) discard the positive set and consider only negative set for summarization, all other setting is similar to $MADS(BiLSTM, \beta = 0.5)$. From table 3 it is evident that $MADS(BiLSTM, \beta = 0.5)$ outperform $MANS(BiLSTM, \beta = 0.5)$. Consequently, it establishes that matching a headline with a summary of a positive and the negative set together is more effective. We further compare $MANS(BiLSTM, \beta = 0.5)$ from table 3 and baseline models from table 2. It is evident that $MANS(BiLSTM, \beta = 0.5)$ outperform both *Feature* and *Encoding* baseline models over NELA dataset. Similarly, $MANS(BiLSTM, \beta = 0.5)$ outperform baseline models FNC (Pomerleau and Rao, 2017), $AHDE$ (Yoon et al., 2019), $HDSF$ (Karimi and Tang, 2019), $FEDS$ (Kim and Ko, 2021b) (Kim and Ko, 2021a), $BiLSTM$, $BERT$ and $RoBERT$ over FNC dataset. From such observations, it is apparent that dual summarization is more effective than considering individual summary of the negative set for the underlying task. But matching a headline with a summary of the only negative set is more effective than summarization-based baseline $FEDS$ (Kim and Ko, 2021b) (Kim and Ko, 2021a) and other state-of-the-art similarity-based baseline models for incongruent news article detection.

Table 4: Comparison of the performances between $MADS(BiLSTM, \beta = 0.5)$ and CDS: Convolution Dual Summary. Here $*$ in $MADS(BiLSTM, \beta = 0.5)$ indicate that $MADS(BiLSTM, \beta = 0.5)$ without convolution summary component and $CDS(BiLSTM, \beta = 0.5)$ is similar to $MADS(BiLSTM, \beta = 0.5)$ without multi-head attention summary component. Results are obtained using attention head $H = 1$ for NELA dataset and $H = 8$ for FNC and ISOT datasets.

	NELA		FNC		ISOT	
Model	Acc	F	Acc	F	Acc	F
MADS(BiLSTM , β = 0.5)	**0.641**	**0.64**	**0.971**	**0.963**	**0.999**	**0.999**
MADS(BiLSTM , β = 0.5)*	0.629	0.605	0.958	0.947	0.998	0.998
CDS(BiLSTM , β = 0.5)	**0.637**	**0.637**	**0.965**	**0.956**	0.998	0.998

4.5 Convolution Versus Multi-head Attention Summary

To study the importance of different summarization components of $MADS$, we compare the performance of $MADS(BiLSTM, \beta = 0.5)$ with $MADS$ without convolution summary component $MADS(BiLSTM, \beta = 0.5)^*$ and CDS (Convolution Dual Summary) differ from $MADS(BiLSTM, \beta = 0.5)$ in considering convolution summary only. From table 4 it is apparent that $MADS$ outperform $MADS$ without convolution summary component $MADS(BiLSTM, \beta = 0.5)^*$ and $CDS(BiLSTM, \beta = 0.5)$. Similarly, superiority of convolution-based summary over multi-head attention-based summary is apparent on comparing the performance of $MADS(BiLSTM, \beta = 0.5)^*$ and $CDS(BiLSTM, \beta = 0.5)$ in table 4.

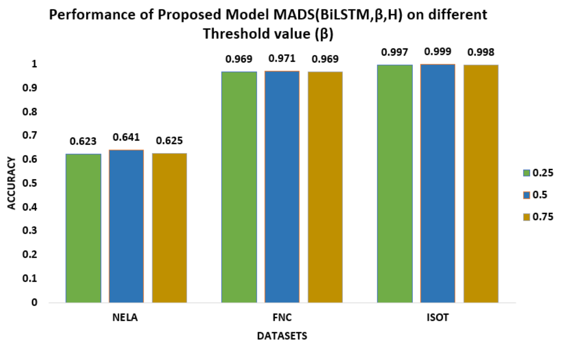

Figure 2: Performance of proposed model $MADS\left(\mathbf{BiLSTM}\ ,\ \boldsymbol{\beta}\ ,\ \mathbf{H}\right)$ on different threshold values $\boldsymbol{\beta}$ over NELA, FNC and ISOT datasets.

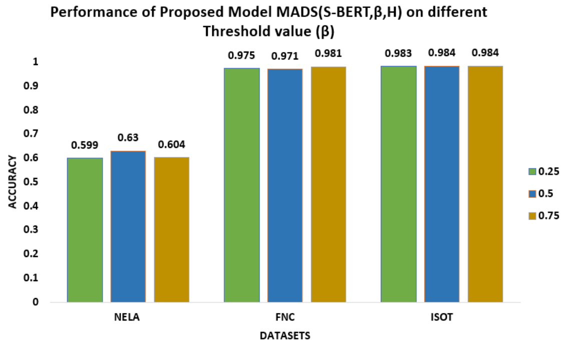

Figure 3: Performance comparison of proposed model $MADS\left(\mathbf{S-BERT}\ ,\ \boldsymbol{\beta}\ ,\ \mathbf{H}\right)$ on different threshold values $\boldsymbol{\beta}$ over NELA, FNC and ISOT datasets.

4.6 Selection of Threshold Value β

The threshold value β is used to split the sentences into positive and negative set. This study considers three different threshold values of β 0.25, 0.5 and 0.75 to produce the results of $MADS(BiLSTM, \beta, H)$ and $MADS(S-BERT, \beta, H)$. From Figure 2 it is apparent that the proposed model $MADS(BiLSTM, \beta, H)$ perform better on threshold value $\beta = 0.5$ across datasets. Similarly, Figure 3 presents the result of $MADS(S-BERT, \beta, H)$ for a different value of β. From Figure 3 it is evident that $MADS(S-BERT, \beta, H)$ performance is superior on $\beta = 0.5$. Hence, $\beta = 0.5$ could be considered as optimal threshold value for both models $MADS(BiLSTM, \beta, H)$ and $MADS(S-BERT, \beta, H)$.

5 Conclusion and Future work

This paper proposed a Multi-head Attention Dual Summarization model, $MADS$, for detecting incongruent news articles of different characteristics. $MADS$ extract two different types of summary, viz. multi-head attention and convolution summary over positive and negative set separately. Subsequently, summaries obtained are matched with headline for incongruent news article detection. It is conclusive from our experimental results that our model $MADS$ is superior in performance to other baseline models across three benchmark datasets. In addition, we conclude that $MADS$ is capable of detecting both incongruent and partially incongruent news articles. This work can be extended to multiple directions in the future. One such direction could be generating topic-aware summarization where the topic of the headline is identified, specific to which the article body is summarized. Generating knowledge-based summarization is another avenue where the summarization is backed by some knowledge bases like Wikipedia etc.

6 Ethics

All the contributions claimed in this paper are original contributions from the authors.

References

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of online fake news using n-gram analysis and machine learning techniques. In *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, pages 127–138. Springer.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Bence Bago, David G Rand, and Gordon Pennycook. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general*, 149(8):1608.

Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining neural, statistical and external features for fake news stance identification. In *Companion Proceedings of the The Web Conference 2018*, pages 1353–1357.

Luís Borges, Bruno Martins, and Pável Calado. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, 11(3):1–26.

Sophie Chesney, Maria Liakata, Massimo Poesio, and Matthew Purver. 2017. Incongruent headlines: Yet another way to mislead your readers. In *Proceedings of the 2017 emnlp workshop: Natural language processing meets journalism*, pages 56–61.

Costanza Conforti, Mohammad Taher Pilehvar, and Nigel Collier. 2018. Towards automatic fake news detection: cross-level stance detection in news articles. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 40–49.

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.

Ullrich KH Ecker, Stephan Lewandowsky, Ee Pin Chang, and Rekha Pillai. 2014. The effects of subtle misinformation in news headlines. *Journal of experimental psychology: applied*, 20(4):323.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Daniel A Effron and Medha Raj. 2020. Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological science*, 31(1):75–87.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*, pages 179–192.

Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, and Felix Caspelherr. 2017. Description of the system developed by team athene in the fnc-1. *Fake News Challenge*.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Benjamin Horne, Sara Khedr, and Sibel Adali. 2018. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Christy Galletta Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. 2021. Emotions: The unexplored fuel of fake news on social media. *Journal of Management Information Systems*, 38(4):1039–1066.

Joonwon Jang, Yoon-Sik Cho, Minju Kim, and Misuk Kim. 2022. Detecting incongruent news headlines with auxiliary textual information. *Expert Systems with Applications*, 199:116866.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. 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 3432–3442, Minneapolis, Minnesota. Association for Computational Linguistics.

Gihwan Kim and Youngjoong Ko. 2021a. Effective fake news detection using graph and summarization techniques. *Pattern Recognition Letters*, 151:135–139.

Gihwan Kim and Youngjoong Ko. 2021b. Graph-based fake news detection using a summarization technique. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3276–3280, Online. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Sujit Kumar, Gaurav Kumar, and Sanasam Ranbir Singh. 2022. Textminor at checkthat! 2022: fake news article detection using robert. *Working Notes of CLEF*.

Minh-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.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Rahul Mishra, Piyush Yadav, Remi Calizzano, and Markus Leippold. 2020. Musem: Detecting incongruent news headlines using mutual attentive semantic matching. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 709–716. IEEE.

Rahul Mishra and Shuo Zhang. 2021. Poshan: Cardinal pos pattern guided attention for news headline incongruence. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1294–1303.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776, New Orleans, Louisiana. Association for Computational Linguistics.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Kunwoo Park, Taegyun Kim, Seunghyun Yoon, Meeyoung Cha, and Kyomin Jung. 2020. Baitwatcher: A lightweight web interface for the detection of incongruent news headlines. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 229–252. Springer.

Dean Pomerleau and Delip Rao. 2017. The fake news challenge: Exploring how artificial intelligence technologies could be leveraged to combat fake news. *Fake News Challenge*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Benjamin Riedel, Isabelle Augenstein, Georgios P Spithourakis, and Sebastian Riedel. 2017. A simple but tough-to-beat baseline for the fake news challenge stance detection task. *arXiv preprint arXiv:1707.03264*.

Julio Rieis, Fabrício de Souza, Pedro Vaz de Melo, Raquel Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 357–366.

Tanik Saikh, Arkadipta De, Asif Ekbal, and Pushpak Bhattacharyya. 2019. A deep learning approach for automatic detection of fake news. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 230–238.

Robiert Sepúlveda-Torres, Marta Vicente, Estela Saquete, Elena Lloret, and Manuel Palomar. 2021. Headlinestancechecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, page 100660.

Percy H Tannenbaum. 1953. The effect of headlines on the interpretation of news stories. *Journalism Quarterly*, 30(2):189–197.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Hermitian co-attention networks for text matching in asymmetrical domains. In *IJCAI*, volume 18, pages 4425–31.

Yariv Tsfati, Hajo G Boomgaarden, Jesper Strömbäck, Rens Vliegenthart, Alyt Damstra, and Elina Lindgren. 2020. Causes and consequences of mainstream media dissemination of fake news: literature review and synthesis. *Annals of the International Communication Association*, 44(2):157–173.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wei Wei and Xiaojun Wan. 2017. Learning to identify ambiguous and misleading news headlines. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4172–4178.

Seunghyun Yoon, Kunwoo Park, Minwoo Lee, Taegyun Kim, Meeyoung Cha, and Kyomin Jung. 2021. Learning to detect incongruence in news headline and body text via a graph neural network. *IEEE Access*, 9:36195–36206.

Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2019. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 791–800.

A Hyperparameter Details

Experimental results presented in this paper are produced with following hyperparameter setting as parented in table 5

Table 5: Present details of hyperparameters used to produce results

Hyperparameters	Values
Epoch	40
Threshold value	0.25, 0.5,0.75
No. of Attention Head	1, 2, 8
Batch Size	50
Embedding dimension	200
Learning rate	0.01
Loss Function	Cross Entropy
memory dimension	100