Recurrent Attention and Multi-channel Convolutional Attention Network for News Bias Detection

Abstract

In the era of information explosion, news proliferates and affects people's minds more and more significantly. Human-operated supervisions on news can't deal with current news' upsurge anymore, and researches on automated news bias detection have become essential and necessary. Previous researches mostly detect news polarization at the level of words and medias, while this paper's view on article-level news bias detection tends to be more intricated and has more application scenarios nowadays. Besides, this paper notices news' unique characteristic and exclusively proposes the Recurrent Attention and Multi-channel Convolutional Attention Network (RAMCA), which is a new neural network architecture suitable for text classification on long articles with hidden information clues. Via ample experiments, this paper compares different network architectures, finds out RAMCA's optimized structure, and proves RAMCA's better performance compared with other similar models. Based on experiments' results and discussions, some inspirations on model optimizations are proposed, including the introduction of Channel Attention Mechanism to the field of Natural Language Processing, the utilization of Multi-channel Network, etc. This paper also points out that the similar tasks based on other languages and scenarios may be the direction for further researches.

1. Introduction

News, as a tool to propagate daily information and celebrities' comments, is rewarded with the vast power of installing particular viewpoints to readers and swaying the general populations' opinions unconsciously.[1][2][16][17] According to previous studies, people's systematic exposure to the news with certain polarization can lead to the ideological segregation in society, and can even directly affect constituencies' voting behaviours.[3][4][5][6][40] Since the advent of Web 2.0 Era, information sources have upsurged and proliferated, which amplifies news bias' influence on readers' minds and imposes stress on manual regulations on the news.[7] Contemporarily, humans obviously aren't capable of dealing with the huge workload, and automated news bias has become necessary. Besides, political climate has become increasingly polarized in some countries currently.[8][9] Thus, it's essential to research on news bias in this paper and make the public aware of the existence of news bias.

This paper attempts to explore the method of automated news bias detection, which can function in many scenarios by annotating news' political tendency. For news medias, it can support the intelligent recommendation algorithms to cater for readers' political preference. For news aggregator applications, based on the news' bias labels, they can bring a more comprehensive and objective horizon on daily news to readers, which alleviates the information cocoon[10][11]. For readers, with the awareness of news bias, they can view the contents more rationally and thoroughly. What's more, news bias detection is also the foundation of many further researches, like news bias remover[12][13][14], recommendation algorithms[15], etc.

Many previous researches have discussed and analyzed news bias from the aspect of journalism[23][24] and NLP[25][26]. Most of them detect news bias at the level of sentences[18], news media[19], social media users[20][21], while the relatively scarce article-level ones are actually always more meaningful in reality[22]. Some article-level researches are carried out, but they only aim at judging the fairness of news, without figuring out the specific political stance that the news conveys[27][28][29]. Besides, in terms of methods, although some researchers[30][31] successfully complete the same task as this paper does, they didn't notice the importance of local feature extraction and failed to consider news' unique characteristic: the bias clues are always subtle and hidden between long negligible text.

The purpose of this paper is to detect news bias at the level of article. In order to extract contextual information both globally and locally, we choose Text-RCNN as the baseline model. To optimize the baseline model and improve the detection accuracy, Attention Mechanism and Multi-channel CNN are introduced into the network architecture, which leads to the proposal of Recurrent Attention and Multi-channel Convolutional Attention Neural Network (RAMCA). Besides, this paper also discusses the effect of some crucial parameters, Attention Mechanism and Multi-channel Convolutional Neural Network on news bias detection.

The rest of this paper is organized as follows: Related work is discussed in Section 2. Then, the architecture and principles of RAMCA is illustrated in detail in Section 3. Section 4 describes the experiments, including the dataset and experimental setup, and Section 5 analyses the results of experiments. Finally, the conclusions and possible further researches for future work are summarized in Section 6.

2. Related work

News bias is a topic worth researching in the field of both journalism and NLP. The awareness of news bias starts from the field of journalism. Groseclose et al. (2005) and Gentzkow et al. (2010) demonstrate that news bias exists by empirical studies[16][23]; De Vreese (2004), Perse (2016) and Reynolds et al. (2002) discuss the cause of news bias[43][44][45]; Dardis et al. (2008)[46], Card et al. (2015)[47], Entman (1993)[48] and Chong et al. (2007)[49] analyze the effect of news bias from the aspect of media framing[32][33]. Since the work of Lin et al. (2006)[50], NLP models have been introduced to detect news bias. Early researches are based on computational linguistics and statistical methods (Greene and Resnik, 2009[51]; Recasens et al., 2013[52]; Lin et al., 2008[53]; Ahmed et al., 2010[54]; Sim et al., 2013[55]); with the rise of deep learning, neural-based approaches have been used to study news bias detection (Iyyer et al., 2014[18]; Preoţiuc-Pietro, 2017[20]; Li and Goldwasser, 2019[56]; Xu et al., 2016[57]; Wachsmuth et al., 2015[58]).

Lai (2015)[36] proposed Text-RCNN model, which consists of Bi-LSTM[37] and Text-CNN[34][35]. Some researches optimized sequential models by introducing Attention Mechanism[41]. Inspired by related conclusions, this paper discusses the effect of typical Attention Mechanism[38] and Headline Attention Mechanism[39]. Besides, while processing the results from different channels, this paper replaces the previous concatenating operation[35] with a method in the field of Computer Visual-Efficient Channel Attention Mechanism (ECA)[42], which optimizes the baseline model.

3. Recurrent Attention and Multi-channel Convolutional Attention

Neural Network

By improving baseline Text-RCNN step by step, this paper proposes the RAMCA (Recurrent Attention and Multi-channel Convolutional Attention) Neural Network. This section will start from the entire architecture of RAMCA and then introduce the principles of Article Encoder, Attention Layer, Multi-channel CNN and Bias Detection respectively, which are all the specific improvements on baseline model.

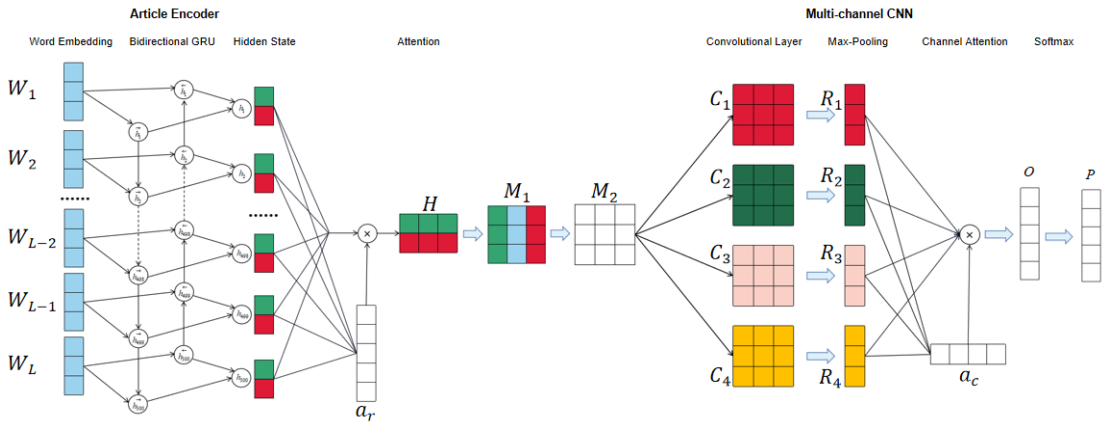

Figure 1: The architecture of Recurrent Attention and Multi-channel Convolutional Attention Network

3.1 Network Architecture

Generally, in consideration of news' unique characteristic: the news articles are commonly prolix and subtle key clues are always hidden between meaningless contexts, the models for bias detection should be equipped with both global and local feature extraction ability. Given an article with L words, RAMCA first uses Article Encoder to process texts via sequential models and extract features globally; then uses Multi-channel CNN to extract features locally and get the final probability distribution P with certain shape.

3.2 Article Encoder

The Article Encoder first uses pre-trained GLOVE model to embed words and get L vectors W_i ($i \in [1,\ L], W_i \in R^{d\times 1}$). The representation matrixes of articles are then input into Bidirectional GRU Layer, to extract features globally from long texts. The Bidirectional GRU can simultaneously process the articles from two directions: the forward $\overrightarrow{GRU}$ reads articles from W_1 to W_L, and the backward $\overleftarrow{GRU}$ reads articles from W_L to W_1. By concatenating the two GRUs' hidden states $[\overrightarrow{h_l}, \overleftarrow{h_l}]$, the i^{th} word can be represented by h_i ($h_i \in R^{2v\times 1}$, $\overrightarrow{h_l} \in R^{v\times 1}$, $\overleftarrow{h_l} \in R^{v\times 1}$).

$$\overrightarrow{h_l} = \overrightarrow{GRU}(W_i) \tag{1}$$

$$\overleftarrow{h_l} = \overleftarrow{GRU}(W_i) \tag{2}$$

$$h_i = [\overrightarrow{h_l}, \overleftarrow{h_l}] \tag{3}$$

Besides, LSTM is another common sequential model to process sentences. Hence, this paper also compares the effect of LSTM and GRU in Article Encoder, and regards LSTM as an alternative

choice for the Recurrent Neural Network.

3.3 Attention Layer

To pay more attention to the key clues hidden in long texts, this paper refers to the previous researches on Attention-based LSTM Networks[38] (Zhou P, 2016) and introduces Attention Mechanism. According to the hidden states h_i, the model calculates the weights vector a_r and assigns weights to certain words. The calculating process is illustrated as following, w_1 is the weight matrix that can changed while training. After multiplying the hidden states by weights vector a_r, the model concatenates the results H and words' representation matrix $[W_1, W_2, \dots, W_L]$ together, and gets M_1. Finally, via linear transformation, M_1 is converted to $M_2 \in R^{128\times 1}$, which will be processed further by Multi-channel CNN in next step.

$$M = \tanh([h_1, h_2, \dots, h_L]); \tag{4}$$

$$a_r = softmax(w_1^T M); \tag{5}$$

$$H = [h_1, h_2, \dots, h_L]a_r; \tag{6}$$

Besides, some researches notice the importance of headlines in terms of news bias detection. Previous papers[39](Gangula R., 2019) assert that a news report's headline can sometimes embody its bias significantly, as headline always recapitulates the entire article's key information. Hence, Headline Attention is proposed to take the correlation between headlines and article contexts into account while assigning weights to certain words. The detailed calculating process of Headline Attention is depicted as following, U is the embedding results of headlines, a_u is the weight vector.

$$a_{u_i} = \frac{e^{{h_i}^T \circ U}}{\sum_i^L e^{{h_i}^T \circ U}}; \tag{7}$$

$$H = [h_1, h_2, \dots, h_L] \circ a_u; \tag{8}$$

This paper regards Headline Attention as an alternative choice for Attention Mechanism, and will compare the effect of two kinds of Attention Mechanisms via experiments in Section 4.

3.4 Multi-channel CNN

Previous researches[36](Chen, 2015) propose that Multi-channel CNN always performs better than Static-CNN, as the negative effect of over-fitting can be reduced to some degree by Multi-channel CNN. Hence, RAMCA optimizes the baseline Text-RCNN model by using Multi-channel CNN to replace the original Static-CNN. Besides, as there's no need to process a single word representation vector by multiple CNN filters separately, the 1-dimension Convolutional Layer and Max-pooling Layer is a more common choice compared with 2-dimension one[59][60]. Hence, the multi-channel CNN part processes the matrix input by 1-dimension Convolutional Layer and 1-dimesnion Max-pooling Layer in order separately to extract features locally. As for different channels, their filters have different region sizes: according to the repetitive experiments, RAMCA has four channels totally, and R_1, R_2, R_3, R_4 are their results. Besides, there are also other papers pointing out that the parameters, like the number of feature maps and region size, have direct and significant effect on the model's performance. This paper will also search for the optimized configurations via experiments in Section 4.

In the field of Computer Visual, Efficient Channel Attention (ECA) is a low-time-costing common method to reduce dimensions. It has three characteristics: (1) it assigns different channels certain weights based on the amount of information they take calculated by Adaptive Pooling Layer; (2) it takes the interactions between different channels while determining channels' weights; (3) it can adjust filters' region sizes on its own according to the specific algorithm. In this paper, RAMCA refers to the ECA method, to integrate the results from different channels. The process is illustrated in detail as Figure 2: Process of Efficient Channel Attention Mechanism.

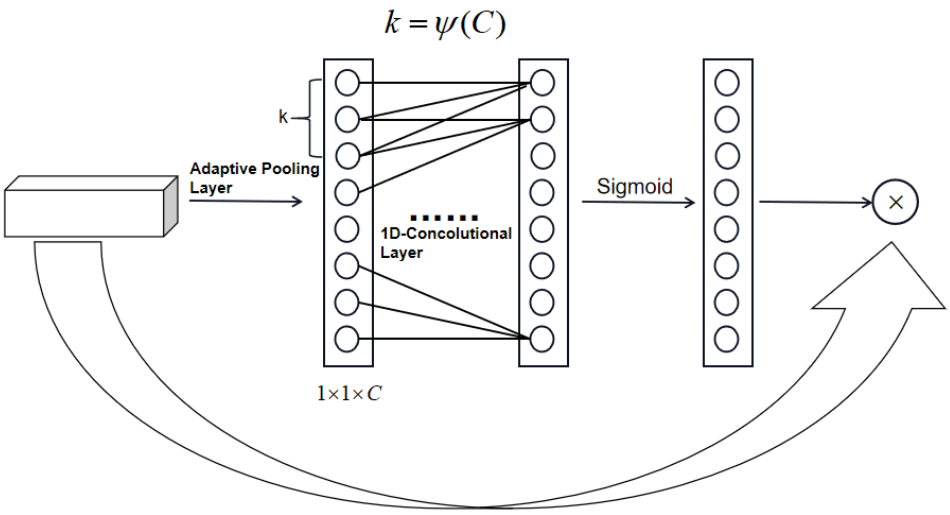

Figure 2: Process of Efficient Channel Attention Mechanism

As k is the region size of 1-dimension convolutional layer, C is the number of channels, b_2 is the bias, γ is the configuration, the self-adjusted algorithm φ can be described as following:

$$k = \varphi(C) = |\frac{\log_2 C}{\gamma} + \frac{b}{\gamma}|_{odd}; \tag{9}$$

Although the multi-channel CNN is introduced into RAMCA, RAMCA doesn't have many channels in total. To avoid some unexpected situation that can lead to the collapse of model training (like the emergence of Null Weight Matrix), experiments maintain the number of channels relatively constant during the training process. Hence, ECA's first two characteristics actually contribute to RAMCA's excellent performance, while the channel number's self-adjusted algorithm doesn't play an essential part. Based on R_1, R_2, R_3 and R_4, ECA can calculate each channel's weight as vector $a_C \in R^{1\times4}$.

Besides, many previous researches[35] directly concatenate the results from different channels for further classification, which can also be deemed as an alternative choice. Hence, this paper also compares the effect of Efficient Channel Attention Mechanism and directly concatenating method as for the integration of results from different channels.

3.5 Bias Detection

Through two fully-connected layer, the results from Multi-channel CNN are transferred to vectors with certain dimensions $O \in R^{5\times1}$ by linear-transformation. Finally, RAMCA's classifier $Softmax$ processes the result vector O and we get the final probability attribution $P \in R^{5\times1}$.

$$P = Softmax(O_i) = \frac{e^{O_i}}{\sum_{i=1}^{5} e^{O_i}}; \quad (10)$$

Based on the probability attribution, the element with the maximum probability is the final detection result.

4. Experiment

Dataset establishment is one of this paper's contributions, and empirical experiment is the main research method in this paper. Hence, this section will first introduce the data source and describe the basic information on the dataset. After that, some layers' hyperparameters and the evaluation indicators will be elaborated specifically.

4.1 Data

In this paper, the experiments depend on the news from Allsides website. Allsides is a authoritative news aggregator platform, which aims at revealing daily events from different aspects and bring a comprehensive and objective horizon to readers. Data from Allsides website has two advantages: (1) all the articles on Allsides website provide annotations of political ideology at the level of articles, which can help to avoid the trivial labelling work; (2) in terms of the same event, there are articles conveying different political tendencies, which prevents the model from binding the certain event and certain news bias together, and improves the model's ability to understand the context's own sake.

News reports are classified into five categories (Left, Lean Left, Center, Lean Right, Right) in total, which is consistent with the contemporary American Political System. The left wing, which is always maintained by the Democrat Party and its constituency, is relatively radical and extreme in terms of voting and making decisions; while the Republican Party advocators mostly support the right wing and tend to be more conservative and reserved.

After removing the meaningless information and deleting some pieces of news with unsuitable length, experiments collected a total of 2949 articles from the website, which are published by 47 powerful medias, including NY Times with left political polarization, WSJ (Wall Street Journal) with center political tendency, CNN (Cable News Network) with left news bias, etc. Collected news shows relatively balanced political tendency distribution, which is illustrated in detail as Table 1: Description on dataset and Figure 3: Political tendency distribution of dataset. Besides, the created dataset is also separated into train dataset, test dataset and validation dataset, for further analysis. Of all 2949 news articles, there are 2057 pieces of training data, 392 articles for test and 500 articles for validation.

Table 1: Description on dataset

Left	Left	250	834
	Lean Left	584	
Center	Center	929	929

Right	Lean Right	290	1186
	Right	896	
Total		2949	

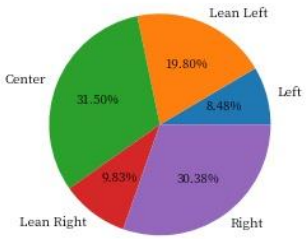

Figure 3: Political tendency distribution of dataset

4.2 Experimental Setup

By observing the average length of the news articles, experiments set 500 as the length of each piece of data ($L = 500$), which means that truncating the articles with more than 500 words and padding the articles whose length are shorter than 500 with zero. In the Embedding Layer, all the words are transferred to vectors with the shape of 200×1 ($W_i \in R^{200\times1}, d = 200$). Each cell in the Bidirectional GRU has 12 layers, which means $H \in R^{24\times500}$. Hence, by concatenating the hidden states and words' representation matrix, the model gets $M_1 \in R^{500\times224}$, and $M_2 \in R^{500\times128}$ after linear transformation. Some other configurations (like region size, number of feature maps) are set according to the further experiments later.

Referring to previous researches and considering the realized situation of this problem, experiments set the CrossEntropyLoss as the loss function for training, and select Adam as the optimizer with parameters $lr = 0.0001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$. Besides, to impede the convergence and improve the training effect, experiments introduce the mini-batch mechanism. More specifically, the mini-batch size is set as 128, and all the following experiments totally have 128 epoches for training.

To compare different models' performance, this paper uses Accuracy, Macro F1, and MAE (Mean Absolute Error) three indicators to evaluate each network's capability. In test dataset, when y_i, $\widehat{y_i}$ means the label value and the prediction value of the i^{th} piece of data separately; for the i^{th} class, t_i means the number of samples with right prediction value and n_i means the total number of samples. As Accuracy equals to the ratio of correctly-predicted samples number to the total samples number, it can illustrate as:

$$Accuracy = \frac{\sum_{i=1}^{5} t_i}{\sum_{i=1}^{5} n_i}; \tag{11}$$

MAE can be shown as:

$$MAE = \frac{1}{\sum_{i=1}^{5} n_i} \sum_{i=1}^{\sum_{i=1}^{5} n_i} |\widehat{y_i} - y_i|; \quad (12)$$

While calculating Macro-F1, each class's indicators $Recall_i$, $Precision_i$ should be calculated first, and Macro-F1 can be illustrated as:

$$Macro - F1 = \frac{\sum_{i=1}^{5} 2\frac{Recall_i \times Precision_i}{Recall_i + Precision_i}}{5}; \quad (13)$$

5. Results and discussion

This paper explores CNN layer's two parameters' effect (Region size, Number of Feature Maps), Impact of Attention Mechanism, Impact of Multi-channel Network via experiments separately and also carries out the Ablation Study to discuss the effectiveness of each RAMCA's improvement on baseline Text-RCNN model. This section will discuss experiments' results and focus on some unexpected findings.

5.1 Effect of Region Size

According to previous researches[35] (Zhang, 2016), region size is a significant contributing factor to the models' final performance. More specifically, Region Size is actually filters' kernel sizes in CNN layers, and can reflect that how many words should be processed together while extracting key information from context to some degree.

To explore the effect of region size and find out the optimized parameter combinations, this paper applies the linear-search method[35]. Hence, experiments firstly use static-CNN architecture to figure out single region sizes' effect on models' performance and estimate the approximate optimized parameters' range as the basis of further experiments on Multi-channel CNN. In this process, experiments consider region sizes of 1, 3, 5, 7, 10, 15, 20, and set the number of feature maps constant as 128.

Table 2: Performance of models with different single region sizes

Region Size	Accuracy	MAE	Macro-F1
1	78.1250%	0.0906	0.7303
3	79.1667%	0.0844	0.6504
5	82.5521%	0.0717	0.7869
7	82.2971%	0.0757	0.7761
10	80.7292%	0.0813	0.7581

15	73.9583%	0.1095	0.5723
20	79.4271%	0.0871	0.7444

Table 2: Performance of models with different single region sizes reveals that a reasonable range for region size might be from 3 to 10. More specifically, when region size is 7, the model performs best; when single region size is 5 or 10, the model also performs relatively well with the accuracy over 80%.

Based on Table 2, this part also explores the effect of combining different filter region sizes. Experiments select different sizes from 3 to 10, and set the number of channels from 2 to 4. Besides, in the process, experiments keep the number of feature maps fixed at 128.

Table 3: Performance of models with different region sizes combinations

Multiple Region Sizes	Accuracy	MAE	Macro-F1
(7, 7)	80.9896%	0.0826	0.765
(7, 7, 7)	82.2917%	0.075	0.7722
(5, 6, 7)	78.3854%	0.0894	0.6341
(6, 7, 8)	82.5521%	0.0752	0.7825
(7, 8, 9)	76.5625%	0.0933	0.6213
(7, 7, 7, 7)	79.4271%	0.0872	0.6493
(3, 4, 5, 6)	79.4271%	0.0859	0.6512
(5, 6, 7, 8)	82.5521%	0.0718	0.7791
(6, 7, 8, 9)	82.8125%	0.0691	0.7918

From Table 3: Performance of models with different region sizes combinations, it's obvious that (7, 7, 7), (6, 7, 8), (5, 6, 7, 8), (6, 7, 8, 9) are some reasonable configurations, which are near the best single region size. More specifically, when the multiple channels' regions sizes are (6, 7, 8, 9), the model performs best, whose accuracy is 84.6354%.

Besides, Table 3 also indicates that more channels in CNN layer don't necessarily lead to better results. To illustrate that, (7, 7, 7) is a better configuration choice than (7, 7) as we expect; but (7, 7, 7, 7) performs significantly worse than (7, 7, 7) with one more channel.

By observing the results further, the parameters adjustment can contribute more when the region

sizes are closer to the best single region size (7). More specifically, the previously-stated multiple region sizes combinations, which are near the best single region size, significantly outperform the model with region sizes (3, 4, 5, 6). Even a Static-CNN model with the region size of 7 performs better than some Multi-channel models. Hence, in other words, the analysis above indicates that the distance between region size combinations and the best single region size is a more contributing factor than other elements. At the same time, the assertion demonstrates the reasonability of linear-search method, which searches the best single region size first to find out the optimized parameters combinations.

5.2 Effect of Feature Map number

According to Zhang (2016), the number of feature maps is another contributing factor that can affect Multi-channel CNN's performance significantly. As it indicates that how many filters are working together to extract features from matrix, this configuration directly decides the shape of each channel's result.

In order to explore the effect of the number of Feature Maps, the experiments keep other configurations constant, and only change the number of Feature Maps for each channel. According to the experiments results in 5.1, the network architecture in this section has 4 channels in total, and the multiple channel regions sizes are 6, 7, 8, and 9 separately. As for Feature Map, experiments consider values $\in \{10, 50, 100, 200, 400, 600, 800, 1000\}$ as the number of Feature Maps.

Table 4: Performance of models with different Feature Map Numbers

Number of Feature Map	Accuracy	MAE	Macro-F1
10	77.3438%	0.0909	0.7239
50	81.25%	0.0757	0.7710
100	82.0313%	0.0729	0.7667
200	79.4271%	0.0833	0.7491
400	82.2917%	0.0726	0.7758
600	83.3333%	0.0656	0.7951
800	83.0729%	0.0681	0.7849

Table 4 shows that each channel's Feature Map number can significantly affect the model's performance. When there are 600 filters in each channel, the model performs best with the accuracy 84.6354%. Besides, when each number's number of feature maps is more than 100, the models' performance has achieved a relatively high level. From 10 to 50, the models' accuracy is

improved more significantly; when the value is over 50, the accuracy fluctuates around 82% and can only be increased slightly for the most time.

5.3 Impact of Attention Mechanism

In the RNN layer, the model aims at assigning higher weights to more important words in long texts by introducing Attention Mechanism. By referring to previous researches, experiments introduce Attention-based Bi-LSTM and Headline-Attention Mechanism into the models' architecture separately and compare the two Attention Mechanisms' function on news bias detection task. Besides, this paper also selects four baseline models as the benchmark to evaluate Attention Mechanisms' effect.

Table 5: Experiments of the Impact of Attention Mechanism

Architecture	Accuracy	MAE	Macro-F1
LSTM	57.0313%	0.1697	0.4772
GRU	58.3333%	0.1673	0.5008
Bi-LSTM	62.2396%	0.1512	0.5191
Bi-GRU	60.4167%	0.1623	0.5203
Bi-LSTM + M-CNN (Multi-channel CNN)	83.3333%	0.0656	0.7951
Bi-LSTM + ATTN + M-CNN(ECA)	84.6345%	0.0636	0.8074
Bi-GRU + ATTN + M-CNN(ECA)	84.1146%	0.0662	0.8014
Bi-LSTM + Headline-ATTN + M-CNN	83.0729%	0.0700	0.7911
Bi-GRU + Headline-ATTN + M-CNN	82.5521%	0.0713	0.7898

From Table 5, it's clear that Zhou's Attention Mechanism[38] improves the models' accuracy significantly by about 5%; when Bidirectional GRU plays as the RNN part, the effect of Zhou's Attention Mechanism is more conspicuous. However, the Headline-Attention Mechanism performs relatively poor in terms of news bias detection problems: the accuracy doesn't change too much after introducing the Headline-Attention Mechanism into the network architecture, and even decreases by 2% when Bidirectional LSTM plays as the RNN part. The news' headlines' characteristics may be a reasonable explanation to this phenomenon: headlines, as the recapitulations of entire news, tend to be objective, and don't include specific information or definite sentimental political tendency; hence, it's reasonable as well as explicable that headlines

can only play a tiny and negligible part while detecting news bias actually. Besides, the difference on languages may be another reason leading to the contradictory between results above and previous researches' conclusions. The proposal of Headline-Attention Mechanism is actually based on Indian news, which is different from this paper's researching basis. As different countries normally have different political systems and reading habits, the language distinction may reward the headlines with different levels of importance.

By the way, it's also noticeable that Bi-LSTM and Bi-GRU outperform the LSTM and GRU. This fact also indicates that Bidirectional RNN plays an essential role in this process, especially while extracting features from long texts.

5.4 Impact of Multi-channel CNN

Previous experiments[34] figured out that Multi-channel models can ease the negative influence of overfitting to some degree, which is the reason why this paper introduces Multi-channel CNN into RAMCA. In this section, there are 4 channels in the model totally, with the region sizes combination of (6, 7, 8, 9); besides, experiments also keep the number of Feature Maps constant at 600 for models' better performance.

Table 6: Experiments on impact of Multi-channel CNN

Architecture	Accuracy	MAE	Macro-F1
Bi-LSTM + Static-CNN	82.2971%	0.0757	0.7761
Bi-LSTM + M-CNN (Multi-channel CNN)	83.3333%	0.0656	0.7951
Bi-LSTM + ATTN + M-CNN(ECA)	84.6345%	0.0636	0.8074
Bi-GRU + ATTN + M-CNN(ECA)	84.1146%	0.0662	0.8014
Bi-LSTM + ATTN + M-CNN(Concat)	83.8542%	0.0678	0.7966
Bi-GRU + ATTN + M-CNN(Concat)	80.2083%	0.0791	0.7579

The results indicate multi-channel's obvious positive effect. As the baseline model with Multi-channel CNN significantly outperforms the model with Static-CNN by more than 7%, Multi-channel networks' improvements on models can be justified. Besides, there are also some examples to support the assertion in previous experiments in Table 5: with similar region sizes, models with Multi-channel architectures normally outperform ones with single channel.

As for the method to integrate results from different channels, Efficient Channel Attention (ECA) Mechanism performs the conventional directly-concatenating method. After getting results from different channels separately, it's necessary to integrate them as a new matrix for further classification. Referring to previous researches, experiments in this section apply ECA and direct concatenation separately, and compare their effects on models' performance. According to Table 6,

ECA shows more significant improvement than simple Concatenation does on models with both LSTM and GRU. Besides, in terms of models' entire architectures, the model with Attention-based Bidirectional GRU, Multi-channel CNN and ECA achieves the highest accuracy and can be regarded as the best architecture for news bias detection task.

5.5 Ablation Study

After exploring the effect of every part of the model and searching for the optimized parameters, this section wants to compare how much each of them can influence the model's performance. Based on the optimized models, experiments separately exclude the Attention Mechanism, Multiple channels part, Bidirectional RNN, RNN layer, and observe each change's outcome.

Table 7: Each part's contribution in RAMCA

Architecture	Accuracy	MAE	Macro-F1
Bi-LSTM + ATTN + M-CNN + ECA	85.6771%	0.0641	0.8200
W/O Bi-RNN	84.1146%	0.0662	0.8021
W/O ATTN	83.8542%	0.0667	0.8000
W/O ECA	83.8542%	0.0678	0.7966
W/O RNN	81.5104%	0.0746	0.7677
W/O M-CNN	82.2971%	0.0757	0.7761

From Table 7, one can see that Attention Mechanism and Multichannel-CNN are both contributing to the models' performance, whose effect and influence are similar. When the Bidirectional LSTM is replaced by LSTM, the model's degeneration is more significant, with a decrease on accuracy by 2%; when the RNN part is eliminated, the accuracy falls almost 5%. As RAMCA's each part can improve the baseline model more or less, and two RAMCA architectures both achieve higher accuracy than the baseline Text-RCNN does, RAMCA's proposal is meaningful in terms of news bias detection problem.

6. Conclusion and Future work

Based on News Bias Detection task, this paper proposes the RAMCA architecture by making improvements on Text-RCNN and uses experiments to prove RAMCA's better capability to process long news texts. According to the analysis above, conclusions in three aspects can be summarized.

In terms of configurations, region size and the number of feature maps are two variables that can influence the models' performance significantly and directly. Linear-search is an effective and

efficient method to determine the ideal and optimized range of flexible parameters.

Discussion on Attention Mechanism is an essential part in this paper. Based on the experiments results, Zhou's Attention Mechanism can help to improve the models' ability to extract key features from long contexts, while Headline-Attention Mechanism doesn't show significant improvement on the model in terms of news bias detection and long texts classification tasks. The results seem conflict to previous researches', but it becomes explicable and explicit while noticing the difference on news' languages. As different countries always have diversified political systems, writing standards and reading habits, it's not sure that if the headlines can always embody the writers' proposal and tendency definitely. Hence, other Attention Mechanism practitioners should pay attention to the realized situation when designing network architectures and selecting methods.

As for the operations on results from different channels, ECA outperforms the concatenation. As ECA can assign different channels with certain weights by evaluating the amount of information each channel carries, the method is just more reasonable and logical than simply concatenating multiple results in one single direction. Consequently, it may be possible to use Channel Attention Mechanism to replace the usual concatenation in further researches, or in other fields.

In the future, RAMCA may be used to analyze news in other languages, and adapted to other countries' political systems. Besides, it's also possible to find more other scenarios to put the model into practice and extend its current applications.

Reference

[1] DellaVigna S, Gentzkow M. Persuasion: empirical evidence[J]. Annu. Rev. Econ., 2010, 2(1): 643-669.

[2] McCombs M, Reynolds A. How the news shapes our civic agenda[M]//Media effects. Routledge, 2009: 17-32.

[3] DellaVigna S, Kaplan E. The Fox News effect: Media bias and voting[J]. The Quarterly Journal of Economics, 2007, 122(3): 1187-1234.

[4] Iyengar S, Hahn K S. Red media, blue media: Evidence of ideological selectivity in media use[J]. Journal of communication, 2009, 59(1): 19-39.

[5] Saez-Trumper D, Castillo C, Lalmas M. Social media news communities: gatekeeping, coverage, and statement bias[C]//Proceedings of the 22nd ACM international conference on Information & Knowledge Management. 2013: 1679-1684.

[6] Dunaway J, Graber D A. Mass media and American politics[M]. Cq Press, 2022.

[7] Baly R, Martino G D S, Glass J, et al. We can detect your bias: Predicting the political ideology of news articles[J]. arXiv preprint arXiv:2010.05338, 2020.

[8] Prior M. Media and political polarization[J]. Annual Review of Political Science, 2013, 16: 101-127.

[9] Pew Research Center. The partisan divide on political values grows even wider[J]. Pew Research Center, 2017.

[10] Ji L. How to crack the information cocoon room under the background of intelligent media[J]. International Journal of Social Science and Education Research, 2020, 3(3): 169-173.

[11] Peng H, Liu C. Breaking the Information Cocoon: When Do People Actively Seek

Conflicting Information?[J]. Proceedings of the Association for Information Science and Technology, 2021, 58(1): 801-803.

[12] Raza S, Reji D J, Ding C. Dbias: Detecting biases and ensuring Fairness in news articles[J]. 2022.

[13] Liu R, Jia C, Wei J, et al. Mitigating political bias in language models through reinforced calibration[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(17): 14857-14866.

[14] Jacobsen G S. An Exploration of Bias and Fairness in Algorithmic Decision-Making Systems[D]. NTNU, 2020.

[15] Zihayat M, Ayanso A, Zhao X, et al. A utility-based news recommendation system[J]. Decision Support Systems, 2019, 117: 14-27.

[16] Gentzkow M, Shapiro J M. What drives media slant? Evidence from US daily newspapers[J]. Econometrica, 2010, 78(1): 35-71.

[17] Matthew Gentzkow and Jesse M Shapiro. 2011. Ideological segregation online and offline. The Quarterly Journal of Economics, 126(4):1799–1839.

[18] Iyyer M, Enns P, Boyd-Graber J, et al. Political ideology detection using recursive neural networks[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 1113-1122.

[19] Baly R, Karadzhov G, An J, et al. What was written vs. who read it: news media profiling using text analysis and social media context[J]. arXiv preprint arXiv:2005.04518, 2020.

[20] Preoţiuc-Pietro D, Liu Y, Hopkins D, et al. Beyond binary labels: political ideology prediction of twitter users[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2017: 729-740.

[21] Darwish K, Stefanov P, Aupetit M, et al. Unsupervised user stance detection on twitter[C]//Proceedings of the International AAAI Conference on Web and Social Media. 2020, 14: 141-152.

[22] Kulkarni V, Ye J, Skiena S, et al. Multi-view models for political ideology detection of news articles[J]. arXiv preprint arXiv:1809.03485, 2018.

[23] Groseclose T, Milyo J. A measure of media bias[J]. The Quarterly Journal of Economics, 2005, 120(4): 1191-1237.

[24] Shanto Iyengar and Kyu S Hahn. 2009. Red media, blue media: Evidence of ideological selectivity in media use. Journal of communication, 59(1):19–39.

[25] Sean Gerrish and David M Blei. 2011. Predicting legislative roll calls from text. In Proceedings of the 28th international conference on machine learning, pages 489–496.

[26] Chen W F, Wachsmuth H, Al Khatib K, et al. Learning to flip the bias of news headlines[C]//Proceedings of the 11th International conference on natural language generation. 2018: 79-88.

[27] Derczynski L, Bontcheva K, Liakata M, et al. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours[J]. arXiv preprint arXiv:1704.05972, 2017.

[28] Thorne J, Vlachos A, Christodoulopoulos C, et al. Fever: a large-scale dataset for fact extraction and verification[J]. arXiv preprint arXiv:1803.05355, 2018.

[29] Nakov P, Barrón-Cedeno A, Elsayed T, et al. Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims[C]//International conference of the cross-language evaluation forum for european languages. Springer, Cham, 2018: 372-387.

[30] Gentzkow M, Shapiro J M. What drives media slant? Evidence from US daily newspapers[J]. Econometrica, 2010, 78(1): 35-71.

[31] Gerrish S M, Blei D M. Predicting legislative roll calls from text[C]//Proceedings of the 28th International Conference on Machine Learning, ICML 2011. 2011.

[32] Tankard Jr J W. The empirical approach to the study of media framing[M]//Framing public life. Routledge, 2001: 111-121.

[33] Matthes J. What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990-2005[J]. Journalism & mass communication quarterly, 2009, 86(2): 349-367.

[34] Chen Y. Convolutional neural network for sentence classification[D]. University of Waterloo, 2015.

[35] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.

[36] Lai S, Xu L, Liu K, et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.

[37] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[38] Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016: 207-212.

[39] Gangula R R R, Duggenpudi S R, Mamidi R. Detecting political bias in news articles using headline attention[C]//Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2019: 77-84.

[40] Stefano DellaVigna and Ethan Kaplan. 2007. The fox news effect: Media bias and voting. The Quarterly Journal of Economics, 122(3):1187–1234.

[41] Zhang L, Zhu G, Mei L, et al. Attention in convolutional LSTM for gesture recognition[J]. Advances in neural information processing systems, 2018, 31.

[42] Wang Q, Wu B, Zhu P, et al. Supplementary material for 'ECA-Net: Efficient channel attention for deep convolutional neural networks[R]. Tech. Rep.

[43] De Vreese C. The effects of strategic news on political cynicism, issue evaluations, and policy support: A two-wave experiment[J]. Mass Communication & Society, 2004, 7(2): 191-214.

[44] Perse E M, Lambe J. Media effects and society[M]. Routledge, 2016.

[45] McCombs M, Reynolds A. News influence on our pictures of the world[M]//Media effects. Routledge, 2002: 11-28.

[46] Dardis F E, Baumgartner F R, Boydstun A E, et al. Media framing of capital punishment and its impact on individuals' cognitive responses[J]. Mass Communication & Society, 2008, 11(2): 115-140.

[47] Card D, Boydstun A, Gross J H, et al. The media frames corpus: Annotations of frames across issues[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). 2015: 438-444.

[48] Entman R M. Framing: Towards clarification of a fractured paradigm[J]. McQuail's reader in mass communication theory, 1993: 390-397.

[49] Chong D, Druckman J N. Framing theory[J]. Annu. Rev. Polit. Sci., 2007, 10: 103-126.
[50] Lin W H, Wilson T, Wiebe J, et al. Which side are you on? Identifying perspectives at the document and sentence levels[C]//Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X). 2006: 109-116.
[51] Greene S, Resnik P. More than words: Syntactic packaging and implicit sentiment[C]//Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics. 2009: 503-511.
[52] Recasens M, Danescu-Niculescu-Mizil C, Jurafsky D. Linguistic models for analyzing and detecting biased language[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 1650-1659.
[53] Lin W H, Xing E, Hauptmann A. A joint topic and perspective model for ideological discourse[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg, 2008: 17-32.
[54] Ahmed A, Xing E. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010: 1140-1150.
[55] Sim Y, Acree B D L, Gross J H, et al. Measuring ideological proportions in political speeches[C]//Proceedings of the 2013 conference on empirical methods in natural language processing. 2013: 91-101.
[56] Li C, Goldwasser D. Encoding social information with graph convolutional networks forpolitical perspective detection in news media[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2594-2604.
[57] Xu J, Chen D, Qiu X, et al. Cached long short-term memory neural networks for document-level sentiment classification[J]. arXiv preprint arXiv:1610.04989, 2016.
[58] Wachsmuth H, Kiesel J, Stein B. Sentiment flow-a general model of web review argumentation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015: 601-611.
[59] Guo B, Zhang C, Liu J, et al. Improving text classification with weighted word embeddings via a multi-channel TextCNN model[J]. Neurocomputing, 2019, 363: 366-374.
[60] Zhang C, Guo R, Ma X, et al. W-TextCNN: A TextCNN model with weighted word embeddings for Chinese address pattern classification[J]. Computers, Environment and Urban Systems, 2022, 95: 101819.