CUCPLUS at CheckThat! 2023: Text Combination and Regularized Adversarial Training for News Media Factuality Evaluation

Notebook for the CheckThat! Lab at CLEF 2023

Chenxin Li[1], Ruijin Xue[1], Chichen Lin[1,*], Weijian Fan[1,*] and Xiao Han[1]

[1]*Communication University of China, Beijing, 100024, China*

Abstract

The increasing usage of social media and digital technology has revolutionized the dissemination of news and information, making it easier to share a vast array of events and information. However, the credibility of news shared on social media continues to be a concern, as it often includes misleading and inaccurate information. Nonetheless, the traditional news media still plays a critical role in publicizing event information. The point of view on how to combat fake news has shifted towards the evaluation of news media credibility. Warning the source of false news increases the possibility of identifying and addressing it early on during the news release and early communication stages. This study describes our performance in the CheckThat! Lab Task 4 Factuality of Reporting of News Media. For the text feature of news, we conducted targeted preprocessing activities and utilized diverse article phrases to determine media credibility levels. Additionally, we adopted RoBERTa and regularized adversarial training to minimize the impact of redundant data and improve model robustness. Our official submission to the English version of CLEF-2023 CheckThat! Lab Task 4 was ranked first with a MAE score of 0.295.

Keywords

Factuality, Media-level, RoBERTa, Regularized Adversarial Training

1. Introduction

News media plays a vital role in disseminating important information to the public, but there is a growing concern regarding the biases of different media outlets when reporting specific events [1]. Therefore, it is crucial to ensure the authenticity of news reports, although research in this area is currently lacking. False information and fake news detection is primarily related to solving this issue. Researchers in the field of natural language processing have attempted various techniques [2], such as stance detection [3], rumor detection [4], and fact-checking organizations such as Snopes, FactCheck and Politifact [5]. However, the fact-checking speed of the methods is far from adequate as they cannot keep up with the speed of information transmission. Consequently, they fail to effectively prevent the early-stage dissemination of false news. Furthermore, delayed fact-checking is often distrusted. Hence, the approach of combating

false news is now centered on its release sources. This involves assessing the credibility of news media and warning them of the presence of false news. With early warning, it is more likely to identify false news during the news release stage and early dissemination, and take appropriate measures.

To continue to focus on media credibility assessments, we attended the CheckThat! Lab at CLEF 2023 [6]. The CheckThat! Lab has held workshop for five tasks. We participate the English version of Task 4 [7] which requires to determine the level of factuality of a document or a news outlet. Our official submission was ranked first with a MAE score of 0.295.

Our main contributions can be summarized as follows:

- We analyzed the performance of the existing PLM models including RoBERTa [8], De-BERTa [9], and Longfomer [10] on this task, and finally completed the task with RoBERTa.
- Our observations reveal that news report titles are usually brief and contain the key semantics, whereas the beginning of the article supplements the title, whereas the ending reflects the attitude of the editor or provides a summary of the article. We conducted an experimental comparison of the various input combinations of articles parts to observe differences in the results.
- Furthermore, we also noted that different media outlets vary in terms of the number of articles, length, and topics of focus, which was also a source of inspiration. So we incorporated R-AT [11] to improve our model's effectiveness.

The remainder of this paper is organized as follows. In Section 2, we provide relevant work relevant to this study. We will then discuss the method in Section 3. Results of the experiments and a detailed discussion are provided in sections 4,5. Finally, we have summarized the studies in Section 6.

2. Related Work

The existing work on media credibility assessment mainly tests the text content of various news media websites to judge the authenticity of their reports [12]. Baly et al. [5] first proposed the media credibility assessment task, because previous work assessed credibility based on a single news, tweet or statement, and here the authors view the text published by news sites as a whole. The author used the ordinal number of age prediction in computer vision regression model for media credibility evaluation and media political bias assessment, and made a lot of experiments to verify the different characteristics of the performance of the results, the author found that the text content is the biggest influence on performance characteristics, using multi-task joint model can improve performance. Indeed, their experimental results on manual fact-checking and classification datasets demonstrate that state-of-the-art text representations can only achieve about 65-71% true prediction accuracy and 70-85% biased prediction accuracy, depending on the dataset.

Multimedia has become an important carrier for all news media to deliver news and information. A report usually includes multiple modal information, and the study of tampering detection in photos and videos has been deepening into [13], so visual features are increasingly added to predict the authenticity [14, 15] of news media reports.
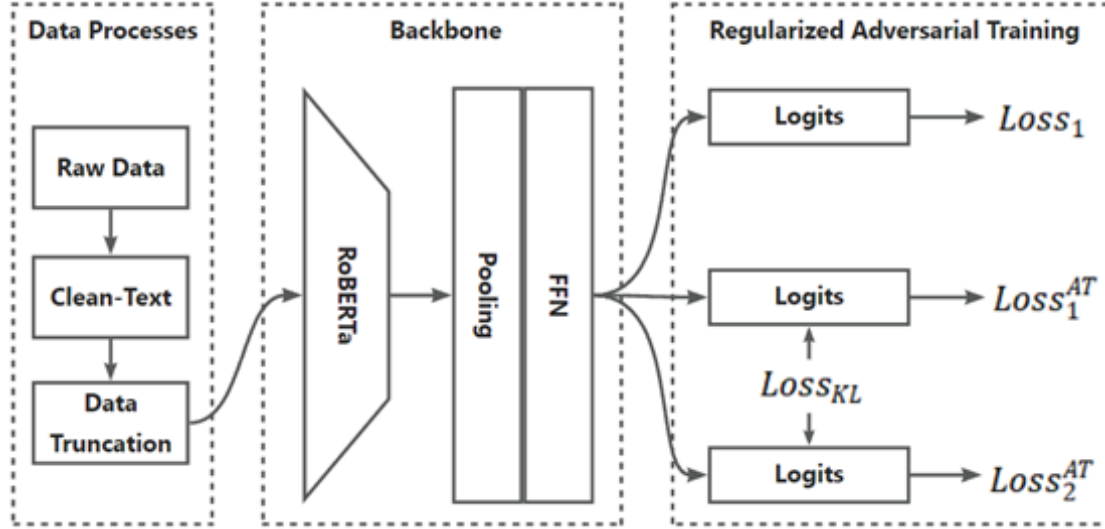
Figure 1: The overall framework of our approach.

In addition to multimodal information such as text and images, it can also detect the relevant features such as the domain name, certificate and hosting properties [16], web design [17] and its linked website [18], so as to estimate the credibility of the news media. Baly et al. [19] further used social context information from the media and explored the impact of news versus social context.

An alternative method used to assess news media credibility is based on news media website similarity. A more reliable indicator of the similarity of news media sites is to examine how much followers of different news media overlap [20]. Panayotov et al. [21] proposed a new model based on graph neural networks that models the audience overlap between media to predict the authenticity and bias of the whole news media. The effect is significantly improved than the previous model.

Evaluating news media credibility can also be based on prior knowledge, Hardalov et al. [22] believe that people believe in manual fact checking and therefore propose to perform automatic fact checking by verifying whether the input statement has been checked by a professional fact examiner and returning an article explaining its decision.

3. Method

The overall framework of our approach is shown in Figure 1. It consists four parts: data preprocess, text combination, RoBERTa and regularization adversarial training.

For Task 4: a level of factuality of a document or a news outlet. We treat it as a classification task. Our input is the collection of the first N articles in the same news outlet (A1, A2, ..., AN). Mosaic of the article that outputs the predicted probability of each confidence level reported(P1, P2, P3).

Firstly, we used the tools to preprocess the articales and remove the noise words. Secondly,

we take some articles as input, set three different input forms and get different input data; then input different forms of text into RoBERTa to output the prediction probability of each credibility level; we also build an adversarial training process using R-AT to help the model fight overfitting and learn more universal features. Finally, we aggregate the prediction results of the models trained by different input forms to obtain the final output results.

3.1. Data Preprocess

Through the observation of the text, we find that some news text only has the title and lacks the content, and the news text contains a large amount of noise data such as URL, special symbols and etc.

For title-only news, we copy the title as its content text.

For noise in news text, we used the clean-text [1] tool to clear line breaks, ASCII codes, URL, email, and punctuation from the title and content, replace all numbers with [number] and convert all uppercase letters to lower case letters.

3.2. Text Combination

For the news text, the author of the news needs to describe the story by the whole article in as concise language as possible, the title is often the most critical semantic information to let the reader get direct understanding of the article. In addition, we found that the beginning and end of the news tend to provide more important semantic information.

The first sentence or the first paragraph of the news often serves as the lead language of the whole article, briefly revealing the content of the full text. The last sentence or the last paragraph of the news captures the overall content of the article and gives the final conclusion, which often contains the main point that the author wants to convey.

Based on the above observations, we used three ways to recombine the text and then input it into the model.

- Only title: We try to use only all the headlines to which the media belongs as input, connecting each title with the separator Token [SEP]:

$$\{[CLS]A_1^{title}[SEP]\ldots\ldots A_N^{title}[SEP]\}, \tag{1}$$

- Head-tail combination: Since RoBERTa can handle up to 512 Tokens in each sequence (multiple news text splicing), we will only take a part of the news content as a representative of the whole story. Following our observations, we used the Head-Tail Truncation approach to process the input news text content and connect with the title using the separator Token [SEP]:

$$\{[CLS]A_1^{title}[SEP]A_1^{head} + A_1^{tail}[SEP]\ldots\ldots[SEP]A_N^{head} + A_N^{tail}[SEP]\}, \tag{2}$$

- Only tail: we found that the text tail truncation way can achieve and close to only title performance, but predict the correct results and only title correct result only less overlap,

---

we think may be tail truncation to obtain the conclusion of the article, so as to better predict some suspected of title fraud, so we extra built using the title and text content end as input:

$$\{[CLS]A_1^{title}[SEP]A_1^{tail}[SEP]\ldots\ldots[SEP]A_N^{title}[SEP]A_N^{tail}[SEP]\}, \qquad (3)$$

where A_i^{title} is the title of each new, A_i^{head} is the truncated text of the starting part of each news content, A_i^{tail} is the truncated text of the end part of each news content, and [CLS] is the special token corresponding to the feature vector used for classification.

3.3. Factuality Assessment

We used the basic language model of RoBERTa to assess news outlet credibility. The predicted score for each confidence level is obtained by passing the output of [CLS] token from RoBERTa to the pooling layer, and then passing through the two linear layers and dropout activated by the GELU to obtain the final output. The sizes of the two linear layers are (1024,128) and (128,3) respectively.

Since we observed that in the training set and validation set, high confidence media accounted for the majority of the overall proportion, mixed credibility and low confidence media accounted for less, so we used the Focal Loss [23] alleviate class imbalance problem, the loss function reduces the weight of a large number of simple samples in training, let the model pay more attention to difficult, misassigned samples. We assigned a value to the α coefficient for each category in Focal Loss, such as a 0.35 α for high confidence.

$$Loss_1 = -\alpha \times (1-p)^\gamma \times log(p), \qquad (4)$$

where p is the predicted probability corresponding to the true credibility level of the sample, α is the weight coefficient corresponding to each category to control the relative importance of the loss of different categories, γ represents the gap between the difficult sample and the simple sample. Its values higher, the loss of the difficult sample more important.

3.4. Regularized Adversarial Training

Although the input text was preprocessed by the data, there is still a large amount of information irrelevant to the credibility of the media reports, so we used adversarial training to enhance the generalization ability of the model and avoid serious overfitting of the model because of learning information irrelevant to the true label.

Adversarial training is a commonly used training strategy to enhance the robustness of the model, but it is easy to reduce the performance of the model due to the inconsistency between training stage and testing stage. Therefore, Shiwen Ni et al. proposed a training strategy to regularize the output of adversarial training through dropout, R-AT, which makes the output probability distribution of different sub-models generated by dropout remain consistent under the same adversarial sample.

Specifically, R-AT dynamically perturbs the input embedding vector to generate adversarial samples and then passes the adversarial samples to two dropout sub-models to generate two

probability distributions. Ultimately, R-AT minimizes the bidirectional KL divergence of the two output distributions as predicted by a regularization model.

We replaced the original cross-entropy loss of R-AT with the Focal Loss, and in each training step, we first performed a forward propagation and backpropagation of a clean sample.

Then following the R-AT setting, we using the rapid gradient method (FGM) proposed by Miyato et al. [24] to perturb the word embedding layer of the model, and then conduct two forward propagation to obtain two confrontation samples, and calculate the corresponding classification loss and KL divergence of the two confrontation samples:

$$Loss_1^{AT} = -\alpha \times (1 - p_1^{AT})^\gamma \times log(p_1^{AT}),\tag{5}$$

$$Loss_2^{AT} = -\alpha \times (1 - p_2^{AT})^\gamma \times log(p_2^{AT}),\tag{6}$$

$$Loss_{KL} = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2}.\tag{7}$$

The loss function of the second backpropagation is finally obtained:

$$Loss_2 = Loss_1 + k \times (Loss_1^{AT} + Loss_2^{AT}) + (1 - k) \times Loss_{KL},\tag{8}$$

where p_i^{AT} is the result of the output of the model after using FGM, D_{KL} is the KL distance between two vectors. We rewrite the original loss function of R-AT, rewriting the original fixed weight parameter as an adaptive hyperparameter k, which is used to adjust the weight of multiple different loss functions during the second forward propagation.

4. Experiments

4.1. Task Description

The task was organized as CLEF 2023 - CheckThat! Lab. In task 4 of CheckThat! Lab, the core idea of task is to determine the level of factuality of a document or a news outlet. The task is defined as follows.

The Task 4 ask to predict the factuality of reporting at the media level, given the URL to a news outlet (e.g., www.cnn.com): low, mixed, and high. This is a English version task.

4.2. Data

CLEF 2023 - CheckThat! Lab provides the data for conducting the factuality level prediction of news outlet, which includes the training set and the validation set to train the model, and later releases the test set to evaluate the model. The evaluation index provided by the competition is MAE (average absolute error). The result is smaller of MAE score, the gap between the prediction label and the real label is smaller, that is, the prediction result of model is more accurate.

$$MAE = \frac{\sum_{i=1}^{n} |\widehat{y_i} - y_i|}{n},\tag{9}$$

Table 1
Label distribution of the datasets

Factuality Level	0	1	2
train	121	233	593
dev	16	32	72

where $\widehat{y_i}$ is the prediction label, y_i is the real label, and n is the number of media.

The official provided data is provided in TSV and JSON format. The TSV file contains three columns of news source, JSON file path and label. The JSON file path points to the JSON file containing the list of articles. A JSON file is named after the media name and contains the news source website, article list, tag text (e. g., high, mixed, low), and tags (digital format, e. g., 2,1,0). The article list contains the title and the content. The distribution of tags in the training and validation sets is shown in Table 1. In both datasets, sites with tag 2, followed by those with tag 1 and those with tag 0 were the least, and the distribution was roughly *1:2:5*.

4.3. Baseline

CLEF 2023 - CheckThat! Lab provided four baseline Model: Mid-label, Majority, Random, and Ngram. However, their baseline is not a learnable model, and to better validate the effectiveness of our pre-and-back truncation approach and R-AT, we additionally designed a baseline method for comparison. The method uses RoBERTa-Large as backbone, adds a pooling layer, two fully connected layers activated by the GeLU function and a softmax layer after the features extracted by backbone. Loss function is Focal Loss for multiple class.

4.4. Experimental Setting

We use the Adam optimizer with a learning rate of 5e-4 and adjusted the learning rate with Cosine Annealing. The Weight decay is 5e-3. In experiments, we use Average Pooling. The loss function is Focal Loss for multiple class with γ = 2, α = [0.1, 0.5, 0.4]. The epoch is 50, the batch size is 8 and the maximum length for the text input is 512.

5. Results and Analysis

5.1. Official Submission

The baseline methods' results for the validation set are shown in Table 2. The Ngram model get the best performance with a MAE score of 0.392.

Our official submission to the English version of CLEF-2023 CheckThat! Lab Task 4 was ranked first with a MAE score of 0.295. The leaderborad is shown as Table 3.

This good performance is partly due to our model fusion, and we use four models to vote. Each model takes the RoBERTa-Large as the backbone and adopts the multi-classification Focal Loss, doing the same data preprocessing for the input of each model. The four models differ

Table 2
Baseline Results on the dev set

Baseline	MAE
Mid-label	0.733
Majority	0.533
Random	0.800
Ngram	0.392

Table 3
Leaderboard

Team	MAE
CUCPLUS	0.295
NLPIR-UNED	0.344
Accenture	0.467
UBCS	0.541
Awakened	0.705
Random	0.943

in the content of input, whether R-AT and whether Mean Pooling. Details of the model are as follows:

- Model 1: The input is the title of the article in the website.
- Model 2: The input is the title of the article in the website, and use R-AT to normalize the confrontation training through dropout.
- Model 3: The input is the title and content of the article in the website, where the content of the first 150 after 100. An Average Pooling layer is connected after the features extracted by RoBERTa-Large.
- Model 4: The input is the title and content of the article in the website, where the content is intercepted after 300. An Average Pooling layer is connected after the features extracted by RoBERTa-Large.

Through the method of model ensemble, the final performance improve, which indicates that different inputs and model structures will make the model have different focuses and biases in predicting media factuality level. Using model voting can more comprehensively predict news outlet credibility, and resulting in higher accuracy.

5.2. Ablation Study

5.2.1. Text Combination

To prove the content of the input can affect the prediction results of the model, as well as to obtain the combinations of inputs with more accurate prediction effects, we conducted the combined experiments on the input content. For all experiment, we adopt RoBERTa-Large as

Table 4
Results of different input

Input	MAE
Title-Only	0.402
F100_B200	0.459
F200_B100	0.451
F100_B250	0.434
F250_B100	0.451
F100_B150	0.443
F150_B100	0.410
F150_B200	0.418
F200_B150	0.418
F50_B250	0.434
F250_B50	0.426
F50_B150	0.434
F150_B50	0.426
F175_B175	0.443
F0_B300	0.402
F300_B0	0.475
F0_B100	0.418
F100_B0	0.451
F0_B200	0.410
F200_B0	0.484
F0_B400	0.426
F400_B0	0.427

backbone of models, and the extracted features of backbone are fed into two linear layers for classification and prediction. We use Focal Loss for multiple class as our Loss function.

The experimental results are shown in Table 4, Titile-Only represents using only the title as input, F100_B300 represents 100 and 300 intercepted news content, others in the same way. According to the experimental results, we can see that the effect of using only the title or intercepting the last 300 is the best, and the effect of intercepting the first 100 and the last 150 and the last 200 is not better. This also confirms that the title has the core semantics of the news, and the beginning and end of the content are often the summary of the news. Therefore, it is feasible for us to only use the title or properly intercept the content of the article, and we can use this method to deal with the long text.

5.2.2. R-AT with Different Weights

We studied the effect of weights in the R-AT loss function, following the recommendations in the R-AT origin paper [10], and we adopted a consistent 0.5 Dropout rate for two outputs, which only adjusted the weight proportion of multiple loss functions computed at the second output.

Since we rewrite the original R-AT loss function as an adaptive R-AT loss function, we only need to adjust k here to change the proportion of the loss function and KL Loss during the second forward propagation in the loss function.

Table 5
R-AT with Different Weights

Weights	MAE
k = 0	0.418
k = 0.01	0.393
k = 0.05	0.418
k = 0.1	0.402
k = 0.3	0.377
k = 0.5	0.410
k = 0.7	0.385
k = 0.9	0.393

The experimental results are shown in Table 5. When k is zero, it represents only using FGM as adversarial training, and with k increased, the proportion of Focal Loss in the second forward spread increased, and correspondingly, the proportion of KL Loss decreased. It can be found that the best effect is achieved when the hyperparameters k is 0.3.

Finally,, we selected the output aggregation of the four models as our final submission. The models we choose are F0_300(k=0), F200_100(k=0.3), Title-Only(k=0.01) and Title-Only(k=0.3), we input the news reports of each media into four models , resulting in a list of four model outputs, and we select the mode in the list as the final prediction for each media.

6. Conclusion and Future Work

In this paper, we propose a model based on regularization adversarial training to evaluate media credibility levels. This model can alleviate the redundancy caused by the large numbers of news in the data set unrelated to the media coverage, effectively avoid the serious overfitting phenomenon caused by learning the information unrelated to the real label, and enhance the generalization ability of the model.

Testing media credibility is a challenging task. In the case of limited data sets, many media coverage fields and wide information, it is difficult to avoid the impact of data set bias and improve the model performance. In future work, we plan to experiment with more pre-process techniques, better model the social environment, adopt a more perfect large-scale pre-training language model and adversarial training strategies to improve the performance of the model in the real environment.

7. Acknowledgments

References

[1] F. Hamborg, K. Donnay, B. Gipp, Automated identification of media bias in news articles: an interdisciplinary literature review, International Journal on Digital Libraries (2018) 1–25.

[2] W. Fan, Y. Wang, Cognition security protection about the mass: A survey of key technologies, Journal of Communication University of China(Science and Technology) 29 (2022). doi:`10.16196/j.cnki.issn.1673-4793.2022.03.009`.

[3] B. Ghanem, P. Rosso, F. Rangel, Stance detection in fake news a combined feature representation, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. URL: https://aclanthology.org/W18-5510. doi:`10.18653/v1/W18-5510`.

[4] A. Zubiaga, M. Liakata, R. Procter, Exploiting context for rumour detection in social media, in: Social Informatics, 2017.

[5] R. Baly, G. Karadzhov, D. Alexandrov, J. Glass, P. Nakov, Predicting factuality of reporting and bias of news media sources, 2018. `arXiv:1810.01765`.

[6] A. Barrón-Cedeño, F. Alam, T. Caselli, G. Da San Martino, T. Elsayed, A. Galassi, F. Haouari, F. Ruggeri, J. M. Struß, R. N. Nandi, G. S. Cheema, D. Azizov, P. Nakov, The clef-2023 checkthat! lab: Checkworthiness, subjectivity, political bias, factuality, and authority, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2023, pp. 506–517.

[7] P. Nakov, F. Alam, G. Da San Martino, M. Hasanain, R. N. Nandi, D. Azizov, P. Panayotov, Overview of the CLEF-2023 CheckThat! lab task 4 on factuality of reporting of news media, in: Working Notes of CLEF 2023–Conference and Labs of the Evaluation Forum, CLEF '2023, Thessaloniki, Greece, 2023.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. `arXiv:1907.11692`.

[9] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, 2021. `arXiv:2006.03654`.

[10] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. `arXiv:2004.05150`.

[11] S. Ni, J. Li, H.-Y. Kao, R-AT: Regularized adversarial training for natural language understanding, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 6427–6440. URL: https://aclanthology.org/2022.findings-emnlp.480.

[12] X. Zhou, R. Zafarani, A survey of fake news, ACM Computing Surveys 53 (2020) 1–40. URL: https://doi.org/10.1145%2F3395046. doi:`10.1145/3395046`.

[13] C. Pasquini, I. Amerini, G. Boato, Media forensics on social media platforms: a survey, EURASIP Journal on Information Security 2021 (2021) 1–19.

[14] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multi-modal framework for fake news detection, in: 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), 2019, pp. 39–47. doi:`10.1109/BigMM.2019.00-44`.

[15] B. Singh, D. Sharma, Predicting image credibility in fake news over social media using

multi-modal approach, Neural Computing and Applications 34 (2021). doi:`10.1007/s00521-021-06086-4`.

[16] A. Hounsel, J. Holland, B. Kaiser, K. Borgolte, N. Feamster, J. Mayer, Identifying disinformation websites using infrastructure features, 2020. `arXiv:2003.07684`.

[17] S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, J. Freire, A topic-agnostic approach for identifying fake news pages, in: Companion Proceedings of The 2019 World Wide Web Conference, ACM, 2019. URL: https://doi.org/10.1145%2F3308560.3316739. doi:`10.1145/3308560.3316739`.

[18] J. D. Fairbanks, N. Knauf, E. B. Georgia, Credibility assessment in the news : Do we need to read, 2018.

[19] R. Baly, G. Karadzhov, J. An, H. Kwak, Y. Dinkov, A. Ali, J. Glass, P. Nakov, What was written vs. who read it: News media profiling using text analysis and social media context, 2020. `arXiv:2005.04518`.

[20] K. Darwish, P. Stefanov, M. Aupetit, P. Nakov, Unsupervised user stance detection on twitter, 2020. `arXiv:1904.02000`.

[21] P. Panayotov, U. Shukla, H. T. Sencar, M. Nabeel, P. Nakov, Greener: Graph neural networks for news media profiling, 2022. `arXiv:2211.05533`.

[22] M. Hardalov, A. Chernyavskiy, I. Koychev, D. Ilvovsky, P. Nakov, Crowdchecked: Detecting previously fact-checked claims in social media, 2022. `arXiv:2210.04447`.

[23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007. doi:`10.1109/ICCV.2017.324`.

[24] T. Miyato, A. M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, 2021. `arXiv:1605.07725`.