Yet at Factify 2022 : Unimodal and Bimodal RoBERTa-based models for Fact Checking

Yan Zhuang[1], Yanru Zhang*[1,2]

[1]University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study, UESTC

Abstract

The development of social networks makes it easier and faster to spread news among people, but the spread of some uncertified news can cause great harm. The 'Factify' task of the 'DE-FACTIFY' workshop aims to solve the multi-modal fact verification problem. In this paper, unimodal and bimodal RoBERTa-based models for fact checking are proposed. The text-only model integrates disturbance on embedding layer, a new loss function and data augmentation by sequential dropout layers into the vanilla RoBERTa. Based on the text-only model, the text-image model changes the text embedding input into the fusion features of texts and images. The experiment results show that after the introduction of fusion features, the model improves slightly, but the best model is still our text-only model. With the best average F1 score of 75.59%, we improve the baseline (53.10%) by 22% and are finally ranked 2nd.

Keywords

Multimodality, Fake News, Text Similarity

1. Introduction

The development of social media technology allows people to express themselves and receive information anytime, anywhere. However, in order to attract the attention of users, some media often publish some eye-catching but unconfirmed news. For example, 77% supporters of Donald Trump, the former US president, held the opinion that 2020 US presidential election was manipulated by "voter fraud" because of the information spread in tweet even though they don't have enough evidence [1]. The situation becomes even worse during the COVID-19 pandemic period. There is an urgent need for a model that can automatically detect whether the claim is fake or not.

Fact checking can be described as, given a claim and some support information, such as documents, images and other claims, we need to judge whether the claim entails the support information. Most claims are evidenced-based so that their veracity can be determined by external knowledge [2]. It is of great importance to take the evidence, or support information into consideration since it helps a lot in reasoning in fact checking [3].

In this paper, text-only model and text-image model are both proposed. The unimodal model based on RoBERTa adds disturbance on embedding layer to boost robustness, creates positive

* Corresponding author
De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada
delecisz@gmail.com (Y. Zhuang); yanruzhang@uestc.edu.cn (Y. Zhang*)

samples through sequential dropout layers to augment data and uses a new loss function for alleviating the difficulty of predicting the image-related labels while the bimodal model based on RoBERTa fuses the text embedding and image features and promotes the interaction between two modalities through the self-attention mechanism in transformer. The experiment results show that both models have good effectiveness. The text-only model performs better and helps us rank 2nd in the multi-modal fact verification task.

The rest of the paper is organised as follows. In section 2, related works about fact verification is provided. Followed by the introduction of the task in section 3. In section 4, the details of our proposed models are discussed. Section 5 contains the experiments results and analysis of different models. Towards the end, section 6 concludes the paper along with future directions.

2. Background

Lots of efforts have been put into fact checking and related research has shifted from single modal with monolingual texts to with multilingual texts to multi-modal with multi-text and multi-image [4, 5, 6, 7, 8, 9]. A multi-level inter-sentence attention model shows competitive performance in 'FEVER' dataset, which consists of 185k samples with a claim and a supporting document [4, 10]. Multilingual transformer-based models, additional metadata and evidence from news stories are combined in multilingual dataset 'X-FACT', which contains 31k short statements in 25 languages.

As for the multimodal situation, [11] shows that augmenting text with image embedding immediately boosts performance. In Event Adversarial Neural Networks (EANN) [12], Text-CNN is adapted to extract textual features and pre-trained VGG-19 [13] architecture with fully connected layer is applied to extract visual features. Besides, a fake news detector and a event discriminator take the concatenated features as input, then predict the label and identify the event label respectively.

Multimodal Variational Autoencoder (MVAE) and EANN have something in common that they use the same visual feature extractor and take the concatenated features for further prediction [14]. However, instead of Text-CNN, MVAE uses recurrent neural networks (RNNs) with bi-directional Long-Short Term Memory (LSTM) cells to extract textutal features. After sampling and reconstructing the concatenation of the both features, the model are trained by optimizing the sum of the reconstruction loss and the Kullback-Leibler (KL) divergence loss.

However, both MAVE and EANN ignore the interactions between the textual and visual features. Vision Transformers (ViT) shows excellent performance in the vision-related tasks [15], based on which, Vision-and-Language Transformer (ViLT) takes fusion of texts and images into consideration, performs faster and competitive and shows excellent performance in vision-language classification tasks such as VQA and MSCOCO [16]. By introducing Mixture-of-Modality-Experts (MOME) Transformer to promote deeper modal interactions, Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts (ULMo) achieves state-of-art results in the VQA and MSCOCO task [17]. The ViLT and ULMo focus more on the interactions between textual and visual features and they may provide a better solution for fact checking.

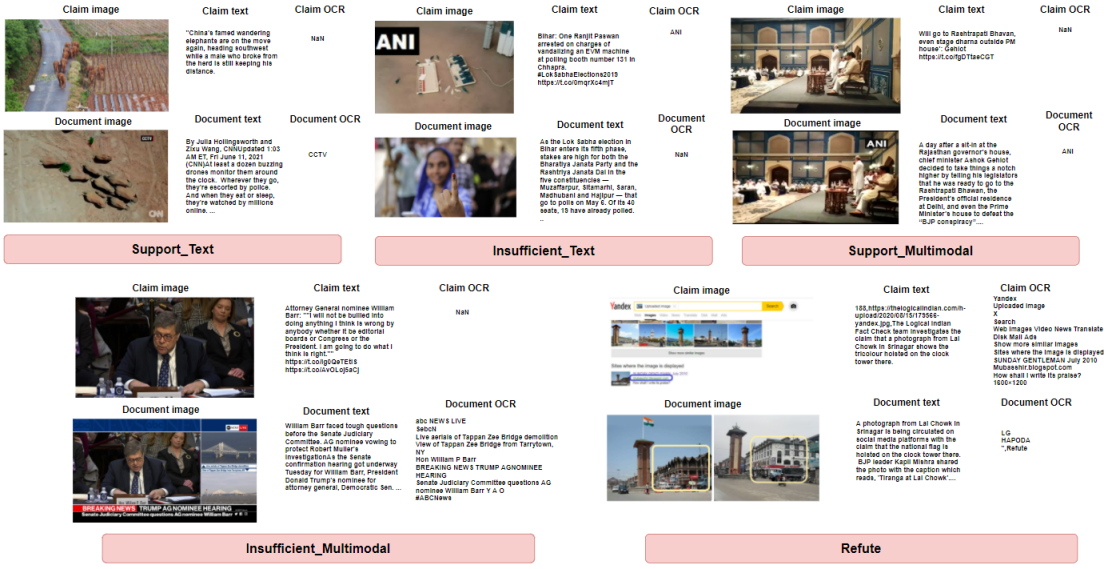

Figure 1: An example of each category in the task dataset [9]. Each sample contains a claim text, claim image, OCR of the claim image and their corresponding document ones. The categories of these samples are judged based on the different entailment mentioned in section 'Task setup'.

3. Task setup

For the task, the dataset contains 50k claims with 100k images [9, 18]. Given a claim text, claim image and OCR of the claim image, we need to predict whether they entail the document ones. According to the different entailment, the claims can be classified into 5 categories:

- Support_Text: the claim text entails the document one but claim image not
- Insufficient_Text: both claim text and image are neither entailed nor refuted by the document ones
- Support_Multimodal: both claim text and image entail document ones
- Insufficient_Multimodal: the claim image entails the document one but claim text not
- Refute: both claim image and text are contradictory with the document ones

In addition, each category accounts for the same percentage with 7k samples for training, 1.5k samples for validation and 1.5k samples for testing. In Figure. 1, a sample of each category is provided.

4. Model

Models can be divided into text-only ones and text-image ones according to the data they use.

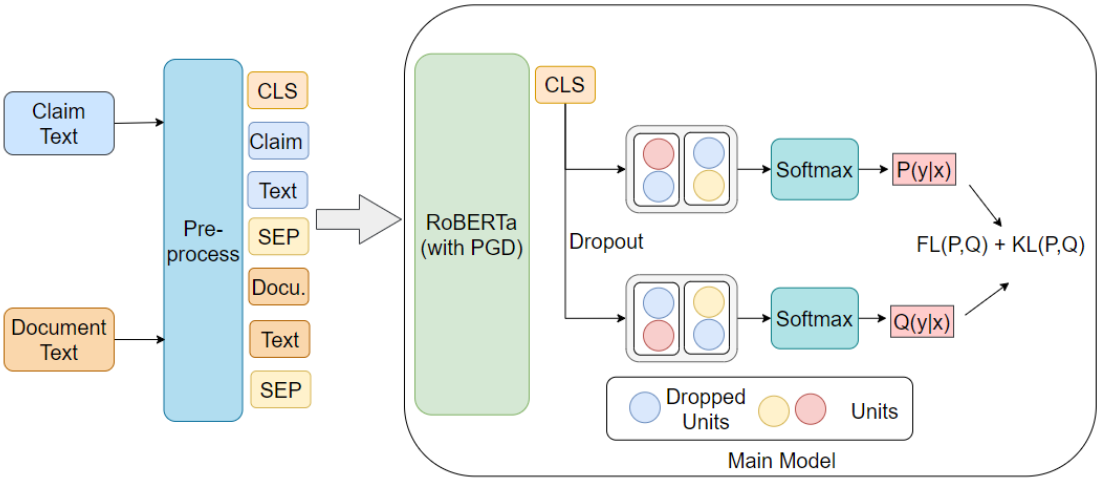

Figure 2: Our text-only model uses RoBERTa as backbone, then the Projected Gradient Descent (PGD) disturbance is put into the word embedding layer [22]. The output of the backbone is later used to create positive and negative samples through R-Drop [23]. The loss function here is defined as the sum of the focal loss and bidirectional Kullback-Leibler (KL) divergence between the two distributions from R-Drop.

4.1. Text Pre-processing

There are lots of meaningless words, URLs and different language characters in claim texts and OCRs, so before feeding the text into the model, the following steps are applied:

- URL removal: There are a lot of URL information in the claims, and the information they contain is worthless and increases the length of the data processed by the model.
- None-English words removal: Many non-English characters are contained in the claims, especially in OCRs, which rarely helps increase the performance.
- Short words removal: There are a lot of spaces, various characters like "`\n`, `aa`" in the original data. Words with less than 3 characters are removed.

Since there is less useful information in the OCRs of the images, many of them are "`NaN`, `ANI`, `BBC`" and the splicing of several words, so the provided OCR data is not used in our model.

4.2. Text-only models

Text-only models treat the task as the sentence pairs similarity problem and solve it by classifying the cosine similarity between the embeddings of the claim and the corresponding document. SentenceBERT [19] is used as for extracting the text embeddings and serves as text-only model baseline in [9]. We use the pre-trained RoBERTa as the backbone and make some modifications [20, 21]. The models structure can be seen in Figure. 2: After removing URL, non-English words and short words, the claim text and document text are fed into the transformer, and here we use vanilla BERT and RoBERTa for comparison. The robustness of the model can be boosted

through introducing disturbance on embedding layer, and we use PGD, which iterates several times to slowly find the optimal perturbation and can be formulated in Equation 1:

$$r_{adv|t+1} = \alpha g_t / \|g_t\|_2 \tag{1}$$

Here g means the input gradient and is defined as equation 2

$$g = \triangledown_x L(\theta, x, y) \tag{2}$$

There are many other adversarial training methods, such as FGM [24] and FreeAT [25]. The former one can only obtain the locally optimal parameters. Although the latter is also a step-by-step iterative search for the optimal disturbance, it is updated based on the gradient and parameters of the previous step, and the parameters found in the current step are suboptimal and do not maximize the Loss.

After we get the last hidden state layer of the CLS in the model, we use a sequencetial network with two dropout layers to generate another CLS layer so as to generate the positive samples, and try to minimize the bidirectional KL divergence between the two CLS layers. The above method is called 'R-Drop' [23]and can be formulated in Equation 3:

$$L_i^{KL} = \frac{1}{2}[KL(Q_\theta(y|x_i)\|P_\theta(y|x_i)) + KL(P_\theta(y|x_i)\|Q_\theta(y|x_i))] \tag{3}$$

In vanilla BERT, the final loss can be computed as the sum of the cross entropy loss. However, since the text-only model only uses text information, it is hard to judge whether the images entail or not. So we add focal loss to alleviate the difficulty of predicting the image-related labels [26]. Here focal loss is defined as Equation 4:

$$L_i^{FL} = -\alpha(1-p)^\gamma log(p) \tag{4}$$

The hyper-parameter α is used to balance the relative importance of positive and negative samples and γ is applied to reduce the weight of easy-to-classify samples, so that the model focuses more on difficult-to-classify samples during training, which satisfies our need. And final loss function of our model is defined as Equation 5:

$$Loss_i = L_i^{FL} + \alpha L_i^{KL} \tag{5}$$

here α denotes the loss weight and we let it equals 4 to compute the loss. The 5 fold cross-validation is also adapted in our model and the averaged logits are used for classification.

4.3. Text-image models

Existing multimodal models focus more on classification tasks with an image and its description [12, 14, 16, 15], and there are relatively few researches dealing with multiple texts and multiple images. The multimodal baseline model provided in [9] computes the cosine similarity between the text embedding derived from the pre-trained SentenceBERT and between the image features derived from the pre-trained ResNet50 respectively [27]. The values of the similarity are seen as the features, as well as the corresponding label as the target, then put into several algorithms

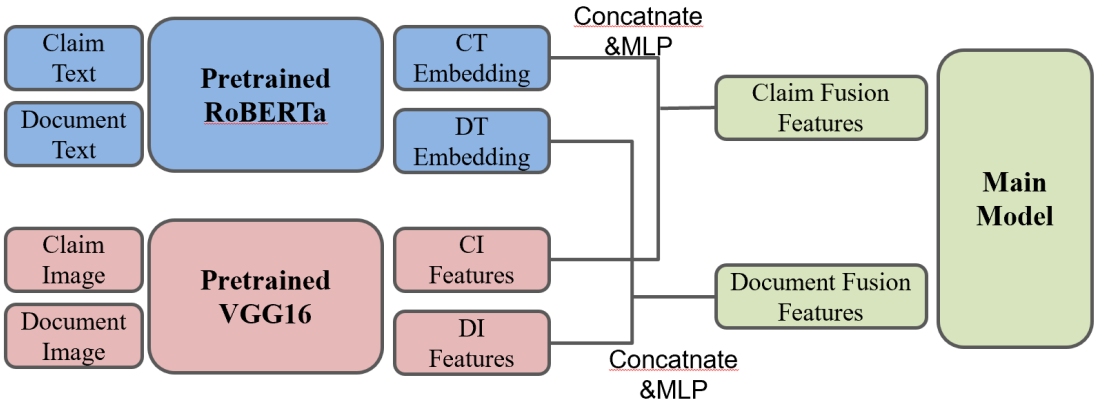

Figure 3: Our text-image model uses Pre-trained RoBERTa to achieve the text embeddings and uses pretrained VGG16 to achieve image features. Then the claim text embedding (CT Embed) and claim image features(CI features), and the document ones are concatenated and fed into Multi-Layer Perception respectively before putting into the main model. The main model is the same as the one in our text-only model.

like Random Forest, Decision Tree, and Logistic Regression. The above baseline model ignores the interactions between different modalities.

Our text-image method shares something in common with baseline that they all use pre-trained model to extract embeddings and features. However, the pre-trained RoBERTa instead of pre-trained SentenceBERT, pre-trained VGG16 instead of pre-trained ResNet50 are applied in our model for their better representation ability. The difference between our text-image model and text-only model is the input. In the former model, the image features are concatenated with the text ones and then put into Multi-Layer Perception for interaction.The whole structure of our text-image model is shown in Figure 4.3.

5. Experiments and evaluations

Here we choose the text-only baseline, pre-trained BERT and RoBERTa as the comparison with our text-only model, use the multimodal baseline and mixed_input RoBERTa without our modifications as a comparison with our text-image model. The mixed_input RoBERTa model takes the fusion of text embeddings and image features, just the way we use in our text_image model, as the model input. The results of all models are classified after averaging all logits obtained from 5-fold cross-validation, except for the two baselines. Besides, all hyperparameters are the same in BERT and RoBERTa models for fair comparison, just as shown in Table 1. The official evaluation for this task is Macro-F1 and the final ranking is based on the weighted average F1 score. The Macro-F1 scores of the models are shown in Table 2. With the best average F1 score of 75.59%, we improve the baseline (53.10%) by 22% and are finally ranked second in this task.

Noting that the models above our text-image model in Table 2 are all text-only models. And

Table 1
The hyperparameters of the BERT and RoBERTa models.

learning_rate	dropout_rate	train_batch_size	adam ϵ
4e-5	0.1	32	1e-8
max_sequence_length	epoch	test_batch_size	seed
128	3	32	42

the column name in the first row of the table except 'Model' is the first few letters of the corresponding label, such as 'Sup_Text' for 'Support_Text', 'Insuffi_Multi' for 'Insufficient_Multimodal'. The figures in the 'Final' column denotes the the weighted average F1 scores of the former 5 categories.

It can be seen that most model perform better in 'Insufficient_Text' and 'Support_Multimodal' label prediction than in 'Support_Text' and 'Insufficient_Multimodal' prediction task for that the judgment basis for the first two labels is that either claim and claim image are both entailed or both not entailed with the document ones. It shows that the information about the interaction between two modalities the models learned is not enough. Besides, all models perform perfectly in predicting the 'Refute' except baselines because it is relatively easy to distinguish texts with the opposite meaning.

The multimodal baseline exceeds text-only baseline over 10% and achieves highest score in 'Support_Text' prediction, but compared with our text-only model, it is over 20% less. The mixed_input RoBERTa model that combines two modalities performs better than the single modality one. Our text-only model shows the best performance among all models and is 1% higher than vanilla text-only RoBERTa. And our text-image model scores higher than mixed_input RoBERTa but does not show competitive performance in image-related label prediction and scores 0.64% less than our text-only model. It is because that the introduction of the image features in RoBERTa decreases the representation ability and results may be the same after interacting different text embeddings and image features. Besides, the difference of the magnitudes may cause bias and variances too. In addition, the ensemble model only ensembles the first 3 models in the Table 2 and performs as well as our text-only model but it costs too much time.

Classifying the multiple texts and images is a tough task for it not only involves the entailment between the texts and texts, images and images, but also between the many texts and images at the same time. Combining the two modalities improves slightly the understanding of the text and image pairs. But better interactions and understanding of the two modalities may further improve the results in future works.

6. Conclusion

In this paper, the unimodal and bimodal RoBERTa-based models are discussed to solve multimodal fact checking task in De-Factify workshop. The major challenge of fact checking task derives from the entailment between multiple texts and images, and existing approaches showed

Table 2
The results of the experiments

Model	Sup_Text	Insuffi_Text	Sup_Multi	Insuffi_Multi	Refute	Final
BERT	58.14%	68.08%	71.06%	64.20%	99.24%	72.14%
RoBERTa	62.01%	70.24%	73.18%	67.96%	99.57%	74.59%
Our text-only model	63.39%	70.85%	74.79%	69.33%	99.60%	**75.59%**
Text_Baseline	-	-	-	-	-	41.33%
mixed_input RoBERTa	62.69%	70.05%	73.90%	68.61%	99.50%	74.95%
Our text-image model	62.90%	70.59%	73.72%	68.41%	99.60%	75.04%
Multimodal_Baseline	**82.68%**	75.47%	74.42%	69.68%	42.35%	53.10%
Ensemble	75.52%	**89.38%**	**82.12%**	**80.81%**	**99.87%**	**75.59%**

unsatisfactory performance. To address the problem, our model integrates the PGD, focal loss and R-Drop into the RoBERTa model, which shows better effectiveness. Besides, our text-image model show better performance compared with the vanilla model by fusing the text embedding and image features, but the effect is still worse than the our text-only model, which helps us stand 2nd in this task. Better multi-modal feature fusion and interaction strategies are conducive to the better solving this challenge.

References

[1] G. Pennycook, D. G. Rand, Examining false beliefs about voter fraud in the wake of the 2020 presidential election, The Harvard Kennedy School Misinformation Review (2021).

[2] S. Shaar, G. D. S. Martino, N. Babulkov, P. Nakov, That is a known lie: Detecting previously fact-checked claims, arXiv preprint arXiv:2005.06058 (2020).

[3] C. Hansen, C. Hansen, L. C. Lima, Automatic fake news detection: Are models learning to reason?, arXiv preprint arXiv:2105.07698 (2021).

[4] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).

[5] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media, arXiv preprint arXiv:1809.01286 (2018).

[6] K. Nakamura, S. Levy, W. Y. Wang, r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, arXiv preprint arXiv:1911.03854 (2019).

[7] J. C. Reis, P. Melo, K. Garimella, J. M. Almeida, D. Eckles, F. Benevenuto, A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 903–908.

[8] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer, 2021, pp. 21–29.

[9] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das,

T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Factify: A multi-modal fact verification dataset, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
[10] C. Kruengkrai, J. Yamagishi, X. Wang, A multi-level attention model for evidence-based fact checking, arXiv preprint arXiv:2106.00950 (2021).
[11] F. Yang, X. Peng, G. Ghosh, R. Shilon, H. Ma, E. Moore, G. Predovic, Exploring deep multimodal fusion of text and photo for hate speech classification, in: Proceedings of the third workshop on abusive language online, 2019, pp. 11–18.
[12] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.
[13] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[14] D. Khattar, J. S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, in: The world wide web conference, 2019, pp. 2915–2921.
[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[16] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, arXiv preprint arXiv:2102.03334 (2021).
[17] W. Wang, H. Bao, L. Dong, F. Wei, Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, arXiv preprint arXiv:2111.02358 (2021).
[18] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Benchmarking multi-modal entailment for fact verification, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
[19] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).
[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[21] 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, arXiv preprint arXiv:1907.11692 (2019).
[22] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).
[23] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu, R-drop: Regularized dropout for neural networks, arXiv preprint arXiv:2106.14448 (2021).
[24] T. Miyato, A. M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, arXiv preprint arXiv:1605.07725 (2016).
[25] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, Advances in Neural Information Processing Systems 32 (2019).
[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[27] P. Kasnesis, R. Heartfield, X. Liang, L. Toumanidis, G. Sakellari, C. Patrikakis, G. Loukas, Transformer-based identification of stochastic information cascades in social networks using text and image similarity, Applied Soft Computing 108 (2021) 107413.