

# Team Noir at Factify 2: Multimodal Fake News Detection with Pre-trained CLIP and Fusion Network

Zhongjian Zhang<sup>1</sup>, Huabin Yang<sup>1</sup>, Chenghao Huang<sup>1</sup> and Yanru Zhang<sup>1,2,\*</sup>

<sup>1</sup>University of Electronic Science and Technology of China

<sup>2</sup>Shenzhen Institute for Advanced Study, UESTC

## Abstract

While social media allows people to easily and quickly obtain information, it also makes low-quality news containing false information spread widely and affect everyone, so it is crucial for fake news detection. Factify 2 is the shared task of the second workshop on multimodal fact checking and hate speech detection at AAI 2023, which aims to classify multimodal datasets containing images, textual claims, reference textual documents and images into five categories. This paper describes the approach we propose for this task. The pre-trained Contrastive Language-Image Pretraining (CLIP) model is used to extract embeddings from texts and features from images. Then these embeddings and features are passed into the fusion network, and calculate the probabilities of each category finally. The ablation study demonstrates the multimodal embeddings and features in fake news detection tasks are more effective than the unimodal embeddings or features. The proposed method reached the best weighted F1 score of 0.745 on the test set and achieved 7th position on the leaderboard.

## Keywords

Fake News Detection, Multimodal, Deep Learning, De-Factify 2@AAAI2023

## 1. Introduction

Social media has changed people's lifestyles in just a short time span. People can easily and quickly obtain different information and meet their social needs. However, it also promotes the widespread spread of misinformation described as "fake news". The term "fake news" gained prominence on social media after the 2016 US presidential election, and its meaning has now evolved into deliberately misleading misinformation that mimics traditional news styles [1]. Nowadays, the spread of false news may destroy the trust in the news ecosystem, damage the reputation of individuals or organizations, cause fear among the general public. Moreover, fake news will also impact politics, economy and other fields, even affect social stability [2]. For example, in the wake of 2020 U.S. Presidential Election, there were a lot of false allegations, many voters who were more focused on election news wrongly believed that election fraud had occurred, and 40% of them insisting that Biden was illegal [3]. Therefore, fake news detection, such as automatic fact-checking and early detection of fake news, can help to eliminate the

---


*De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAI 2023. 2023 Washington, DC, USA*

\*Corresponding author.

✉ a2665446972@gmail.com (Z. Zhang); 202121080420@std.uestc.edu.cn (H. Yang); zydhjh4593@gmail.com (C. Huang); yanruzhang@uestc.edu.cn (Y. Zhang)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

possible adverse effects in advance.

Fake news detection by automatic fact verification is to give a claim and some reference support information and utilize the automatic method to judge whether the claim entails reference. The past decade has seen significant advances in computer vision and natural language processing with the development of artificial intelligence and deep learning. Multimodal tasks, such as text-image generation [4] and visual dialogue [5], are increasingly becoming the focus of researchers. There are some datasets [6, 7, 8] and related work [9, 10, 11] on textual fact verification, but less work on multimodal or cross-modal fact verification. The shared task Factify 2 aims to promote this work by introducing a new multimodal fake news dataset [12] and proposing a baseline for researchers.

This paper proposes a model combining the pre-trained model and fusion network for fake news detection. Pre-trained CLIP [13] is used to extract embeddings from the text of claims and documents and features from corresponding images. The fusion network consists of the extraction module, multi-head attention module, and multi-layer perceptron (MLP). After passing the obtained embeddings and features through the fusion network, the probability of each category is calculated finally. Our approach ended up in 7th position on the Factify 2 task with 0.745 of weighted F1 score. The overview of the task can be found here [14]. Moreover, we conducted a comparison experiment in the ablation study using only text embeddings or image features. The F1 score of multimodal embeddings and features is over 0.1 higher than unimodal embeddings or features, which proves that multimodal information is more effective in fake news detection.

The rest of the paper is organized as follows. Section 2 presents the related work of fake news detection. The Factify 2 dataset used for the task and the pre-processing method is described in Section 3. Section 4 introduces our proposed method. The experiment and results are shown in Section 5. And we conclude the paper and provide some future directions at last.

## 2. Related Work

Fake news detection was completely based on unimodal text in the early stage, and multimodal false news detection has been developed in recent years. Researchers interested in promoting this field have published new datasets [15, 16] and conducted some studies, studies on multimodal fake news detection still need to be further promoted. The detection method at present can be divided into three main categories.

The first is to detect claims to determine authenticity. Multimodal Variational Autoencoder (MVAE) uses VAE to extract the latent code, which fuses the features of images and texts and then passes the latent code through the detector to classify the fake news [17]. [18] proposes Attention-based Multimodal Factorized Bilinear Pooling (AFMB) for multimodal Fake News Detection, Stacked BiLSTM and Multi-level CNN-RNN are used for textual and visual feature extraction respectively, FMB module combines textual and visual features into fusion representation and passes through the MLP for fake news classification. However, MVAE, AFMB, and other works [19, 20, 21] are designed to detect the veracity of claims directly on multimodal text and images, fail to detect fake news by comparing claims with reference documents.

The second detection method is according to the dissemination information of the claim.

Judging only by the content of the claim is often insufficient. The credibility of the user or the propagation path of the claim can also be used for fake news detection. A detection approach is proposed based on features extracted from the user profile and news content, Long Short Term Memory (LSTM) is used for fake news classification, which outperforms K-NearestNeighbor (KNN), Support Vector Machine (SVM) and other machine learning classification methods in detection performance [22]. In [23], the correlation among multiple news, such as time, content and source, is used to build the News Detection Graph (NDG), which is analyzed by the Heterogenous Deep Convolutional Network (HDGCN) to detect fake news.

Fake news can also be detected by comparing the external reference containing objective facts with the claim. An end-to-end graph neural model called CompareNet is proposed in [24], which compares the news with the knowledge base (KB) through entities for fake news detection, contextual entity extracted by heterogeneous graph convolution network and KB-based entity representations are fed into the entity comparison network to capture the consistency between the news and KB. To consider the knowledge-level relationships among news entities in fake news detection, [25] propose a Knowledge-aware Attention Network (KAN), which identifies entity mentions in the news and aligns them with the entities in the knowledge graph to provide complementary information as external knowledge.

Factify is a shared task in the De-Factify workshop at AAAI 2022, which is modeled as a multi-modal entailment task and aims to realize automatic fake news detection [26]. Several researchers have proposed models in this task to solve this problem. Most of them extract text embeddings based on BERT [27] or its variants, while image feature extractors are more diversified, such as VGG-16 or ResNet-50. The weighted F1 of the best results among participants reaches 0.768. This shared task also shows that constructing a model that performs well in all categories of multimodal fake news detection is a challenge that still requires ongoing research.

### 3. Dataset and Pre-processing

#### 3.1. Dataset Description

Factify 2 is a multimodal fake news detection dataset consisting of 50K data samples [12]. Each sample contains seven descriptions: 'Claim', 'Claim\_image', 'Document', 'Document\_image', 'Category', 'Claim\_OCR' and 'Document\_OCR', where 'Claim' and 'Document' are the text of the given claim and the given reference respectively, 'Claim\_image' and 'Document\_image' are the images corresponding to given claim and given reference. 'Claim\_OCR' and 'Document\_OCR' are the OCRs obtained from the corresponding images using the Google Cloud Vision API. 'Category' is the label of sample, there are five types according to other descriptions:

- **Support\_Multimodal:** both the claim text and image are similar of the document;
- **Support\_Text:** the claim text is similar or entailed but images of the document and claim are not similar;
- **Insufficient\_Multimodal:** the claim text is neither supported nor refuted by the document but images are similar to the document;
- **Insufficient\_Text:** both text and images of claim are neither supported nor refuted by the document;

**Table 1**  
Word distribution statistics.

Train Set		Validation Set		Test Set	
Word	Frequency	Word	Frequency	Word	Frequency
cases	7612	cases	1592	cases	1442
covid19	6028	covid19	1177	covid19	1125
minister	4559	minister	903	minister	840
delhi	4020	delhi	833	delhi	834
today	3668	today	769	new	785
new	3580	new	762	today	637
updates	3558	updates	712	total	612
state	3041	state	686	president	514
total	2959	total	615	deaths	487
deaths	2872	president	604	state	462
president	2680	deaths	579	updates	423

- **Refute:** the images and/or text from the claim and document are contradictory.

Each category has the same number of samples (10000), and the dataset is divided into training set, validation set and test set at a ratio of 70:15:15. Only the labels of the training set and validation set are public.

### 3.2. Data Pre-processing

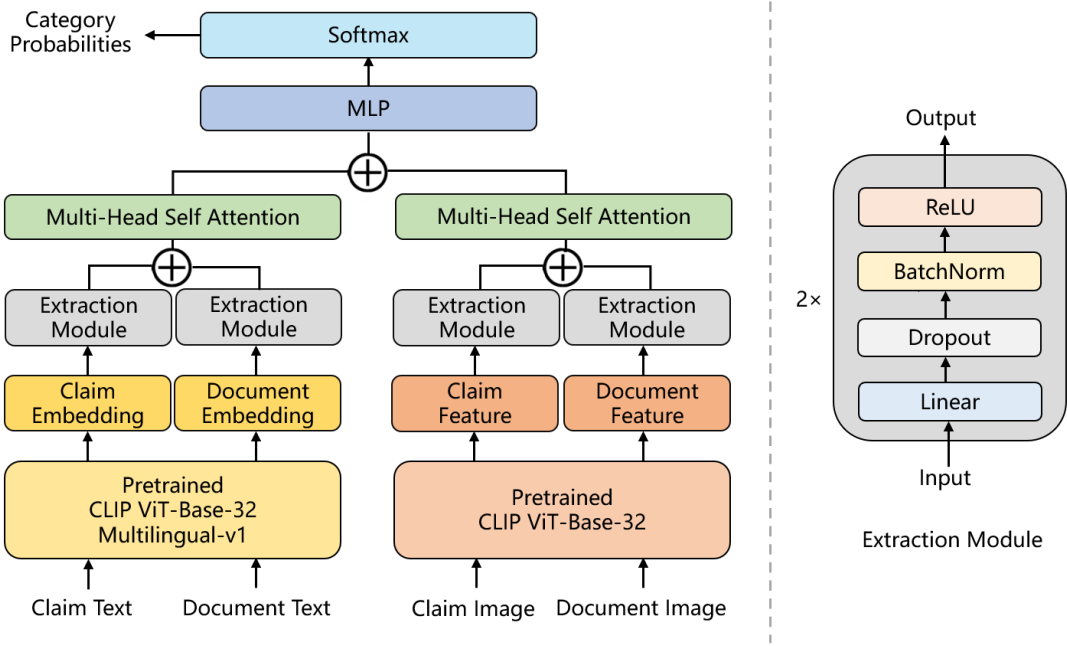
For the text data in the dataset that contains a lot of useless stop words, punctuations, and URLs, the following processing is applied:

- **URLs removal:** There is little or no semantic information in the URLs. Removing these URLs does not change the meaning of the text and reduces the length of the text data to make it easier for the model to extract embeddings from the text.
- **Stop words and punctuations removal:** Stop words and punctuation removal: Similarly, stop words that do not affect text semantics and punctuations are also removed from text data.

After processing, most of the text data is English, while a small part is Hindi. Moreover, the OCRs obtained from the images were not used for model training because many of the data were numbers or incoherent phrases. The images in the dataset are resized to 224 and normalized, then fed into the pre-trained vision model to obtain the features.

### 3.3. Data Analysis

We count the word frequency distribution of the 'claim' text data in the dataset. After removing the conjunctions such as 'the', 'of', the words most frequently appeared in the dataset is shown in the Table 1. It can be seen that the dataset is related to present affairs, policy and government, and the dataset is evenly divided. The ratio of word frequency is almost the same in each dataset.



**Figure 1:** The framework of our proposed model. The pre-trained CLIP model extracts embeddings from texts and features from images. Then the unimodal embeddings and features are fed into the extraction module and concatenated together to obtain the text embeddings and image features. After processing these embeddings and features by the multi-head attention module, they are concatenated together to form fusion features and put into the MLP. Finally, Softmax is used to calculate the probability of each category.

## 4. Methodology

In the previous Factify task, the text embeddings were all extracted from pre-trained models based on BERT, while the image features were extracted from various visual pre-trained models. These pre-trained models are trained independently, and the extracted text embeddings and image features are associated in the later training process.

We notice that the result of fake news detection in Factify is judged according to the relationship between text pairs, image pairs and text-image pairs. Suppose the pre-trained model is trained by text-image pairs, the similarity of matched or similar text or image will be higher. We want to investigate whether such a strategy would improve performance in fake news detection. Therefore, the pre-trained CLIP model is used in our framework to extract embeddings from texts and features from images.

CLIP, trained by contrastive learning with 400 million (image, text) pairs, has extremely powerful feature extraction capabilities, its zero-shot transfer performance is competitive with prior task-specific supervised models. It is widely used in many downstream tasks and achieves excellent performance. We use CLIP-ViT-B-32 to extract image features from images in the dataset. For the text data containing English and Hindi, we use the pre-trained CLIP-ViT-B-32-

multilingual v1 as the text extractor to obtain the text embeddings. The text extractor is trained by multilingual knowledge distillation [28] and supports 50+ languages, where clip-ViT-B-32 is the teacher model, and a multilingual DistilBERT model [29] is trained as the student model.

Then the embeddings and features extracted from the pre-trained CLIP model are fed into the fusion network, which consists of the extraction module, the multi-head self attention module and the MLP.

The extraction module is used to map the data to a low dimensional space to extract deeper features through the linear transformation. It comprises two small modules stacked with the same structure, each of which reduces the dimension of the input by half. Each small module first passes the input through the linear layer, then makes Dropout and BatchNorm, and finally activates it using the ReLU function. Both text embeddings and image features will be processed by extraction modules with the same structure and different parameters.

After the extraction module is applied to the embeddings and features, the newly obtained claim and document text embeddings are concatenated and fed into the multi-head self attention module to obtain the new text embeddings. The same steps are applied to the claim and document image features to get deeper image features. In short, the attention function maps a query and a set of key-value pairs to an output, and the multi-head attention mechanism repeats this calculation multiple times to allow the model to jointly attend to information from different representation subspaces of the data [30]. In self attention, the query, key and value are identical, which is the input to the multi-head self attention module. The structure of multi-head attention module is the same as that of the Transformer, but we change the dimensions of input and output and the number of heads.

The processed text embeddings and image features from different modals are concatenated together as fusion features and fed into the MLP. MLP changes the output size to 5 by reducing the input fusion features through the linear layer three times and the ReLU activation function twice. Finally, the Softmax layer calculates the probabilities of different categories. The framework of our model is shown in Figure 1.

## 5. Experiments and Results

### 5.1. Implement Detail

In the training process of the model, Adam optimizer with learning rate of  $5e-4$  was used to optimize the parameters. The pre-trained model and fusion network are trained 50 epochs together, the batch size is 32, and the dropout rate in the extraction module is set to 0.2. The dimensions in multi-head attention module is set as 256, the number of heads is 4.

The cross-entropy loss function with label smoothing is used in training. Label smoothing replaces the traditional label vector with the updated label vector, it can be expressed as:

$$y_i = \begin{cases} 1 - \epsilon, & i = \text{Target} \\ \epsilon / (N - 1), & i \neq \text{Target} \end{cases} \quad (1)$$

where  $y_i$  is the updated label, and  $\epsilon$  is a small constant, which is set as 0.1 in training, and  $N$  is the total number of categories. Label smoothing can improve model generalization ability and restrain overfitting effectively.

**Table 2**

The results on the test set

Model	Support_Multimodal	Support_Text	Insufficient_Multimodal	Insufficient_Text	Refute	Final
Text-only	0.7874	0.7363	0.5941	0.7019	0.9916	0.6357
Image-only	0.7981	0.6957	0.7829	0.7130	0.8523	0.6275
w/o multilingual	<b>0.8735</b>	0.7363	0.7898	0.7526	0.9785	0.7144
w/o MHA	0.8558	0.7654	0.8034	0.7820	0.9946	0.7339
Proposed	0.8726	<b>0.7710</b>	<b>0.8156</b>	0.7849	<b>0.9970</b>	<b>0.7452</b>
Baseline	0.8272	0.5000	0.7593	<b>0.8024</b>	0.9882	0.6499

Exponential Moving Average (EMA) is also used to improve evaluation performance and increase robustness. The parameter weights of EMA are applied in model evaluation, which is equivalent to the weighted average of the model parameters in training.

## 5.2. Ablation Study

In addition to training the proposed model to evaluate performance on the test set, we also make some modifications to the proposed model as the ablation study to verify the impact of different components on the model performance.

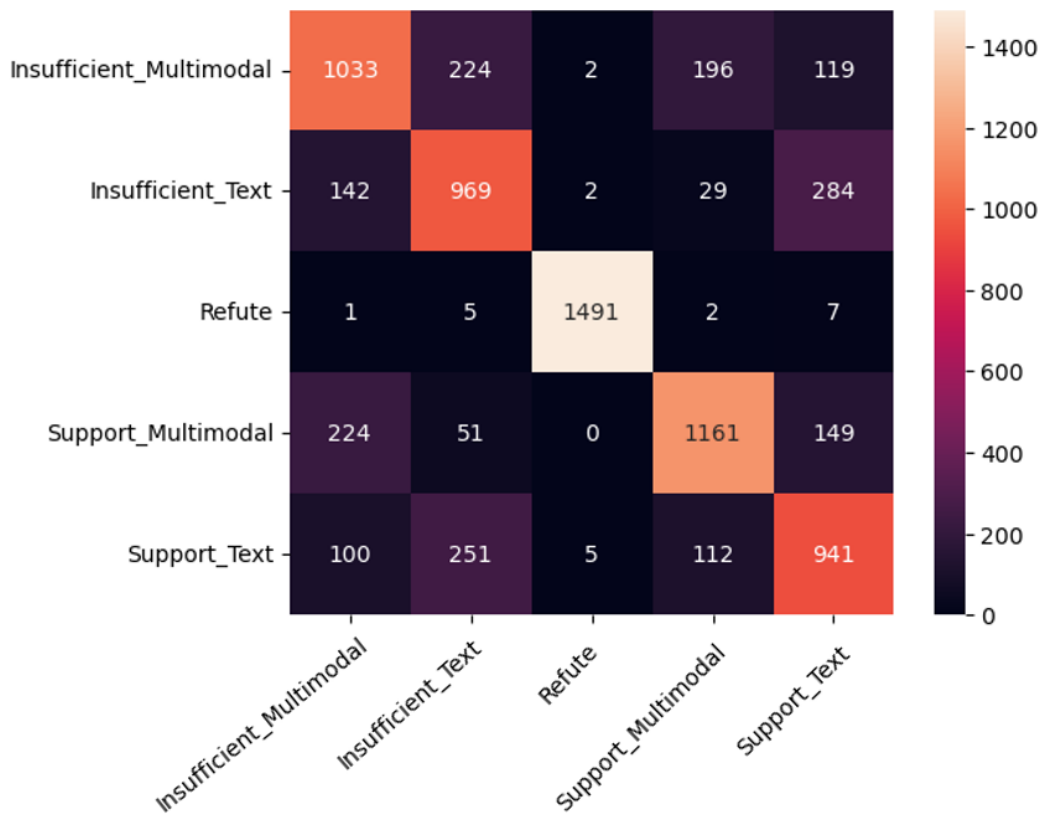
First, we want to analyze the role of different modal input, so the proposed model is modified to use only texts as input or only images as input. Similarly, the text-only or image-only model uses the pre-trained CLIP model to extract embeddings or features and feed them into the extraction module and then the multi-head attention module. The probability is calculated directly through the MLP using Softmax without the fusion step.

To investigate the influence of multilingual texts contained in 'claim' and 'document' in the dataset on the model performance, we use CLIP-VIT-32-B's text extractor instead of CLIP-VIT-32-B multilingual lingual v1 after removing these multilingual texts during data pre-processing. In addition, a model without multi-head attention module is trained to explore the effect of multi-headed attention modules on performance.

## 5.3. Result

The results of the proposed model and other models for comparison on the test set are shown in Table 2. Noteing that 'text-only' and 'image-only' represents the model with text and images as input only, 'w/o multilingual' means the text extractor is the vanilla CLIP-VIT-B-32, and 'w/o MHA' represents the model without the multi-head attention module. 'Proposed' is the model introduced in Section 4, and 'Baseline' was proposed by the workshop organizer. The performance of the model under each category is measured by category-wise F1, and 'Final' denotes the weighted average F1 of five categories.

The performance of the unimodal model with only text or only image as input is close, the text-only model performs better in 'Refute', whose F1 is 0.1393 higher than image-only model, while the F1 of image-only model in 'Insufficient\_Multimodal' is 0.1915 higher than that of



**Figure 2:** The confusion matrix of result on the test set.

text-only model. We infer that this is because 'Insufficient\_Multimodal' is related to whether images are entailed or not, while 'Refute' can be judged as long as texts are contradictory.

The performance of text extractor without multilingual information is much better than the unimodal model, even slightly exceeding the proposed model on 'Support\_Multimodal'. However, it is worse than the proposed model in other categories, with the weighted f1 0.0308 lower, which proves that preserving multilingual text in data can slightly improve the model's performance in fake news detection. Moreover, by comparing the results before and after removing the multi-head attention module from the model, it can be seen that adding multi-head self attention mechanisms does improve the model performance to some extent.

For the proposed model with both texts and images as input, the weighted average F1 is 0.1 higher than that of the text-only model and image-only model, especially on 'Support\_Multimodal' and 'Insufficient\_Text'. It shows that combining text embeddings and image features can improve the model performance and further proves that multimodal data is more effective than unimodal data in fake news detection.

The confusion matrix of the result on the test set is shown in Figure 2. In addition to the 'Refute' category, due to the large difference between 'Insufficient\_Text' and 'Support\_Multimodal',





Figure 3: Examples of model misclassification.

they are less likely to classify to each other. However, other categories may be classified into any category except 'Refute', indicating that the model still needs to be optimized to enhance the ability to distinguish text embeddings and image features to minimize the probability of misclassification and improve accuracy.

Some examples of model misclassification are shown in the Figure 3. Besides the fact that the 'Refute' category data is collected from a different source than the other classes, the main reason for misclassification of all categories except 'Refute' is that some text data is too long. The data used to train the model will be truncated, resulting in a loss of semantic information and subsequent misclassification.

## 6. Conclusion

In this paper, we propose a model for fake news detection using the pre-trained CLIP model and a fusion network. The proposed model achieves 7th position in the final leaderboard for the shared task 'Factify 2'. Moreover, comparative experiments demonstrate the effectiveness of multimodal data in fake news detection. Further work can be carried out in the following aspects: 1) Explore more effective methods to extract embeddings and features from the data, such as using other pre-training models as extractors; 2) Optimize the network model structure

and use better strategies to integrate different modal data; 3) Integrate the results obtained from different models to get a more robust result; 4) Transfer the proposed model to other fake news detection data sets and optimize it to improve its performance on this task.

## References

- [1] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, *Journal of economic perspectives* 31 (2017) 211–36.
- [2] G. Di Domenico, J. Sit, A. Ishizaka, D. Nunan, Fake news, social media and marketing: A systematic review, *Journal of Business Research* 124 (2021) 329–341.
- [3] 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).
- [4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, *arXiv preprint arXiv:2204.06125* (2022).
- [5] V. Murahari, D. Batra, D. Parikh, A. Das, Large-scale pretraining for visual dialog: A simple state-of-the-art baseline, in: *European Conference on Computer Vision*, Springer, 2020, pp. 336–352.
- [6] W. Y. Wang, ” liar, liar pants on fire”: A new benchmark dataset for fake news detection, *arXiv preprint arXiv:1705.00648* (2017).
- [7] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, *arXiv preprint arXiv:1803.05355* (2018).
- [8] A. Gupta, V. Srikumar, X-factor: A new benchmark dataset for multilingual fact checking, *arXiv preprint arXiv:2106.09248* (2021).
- [9] L. Pan, W. Chen, W. Xiong, M.-Y. Kan, W. Y. Wang, Zero-shot fact verification by claim generation, *arXiv preprint arXiv:2105.14682* (2021).
- [10] N. Vedula, S. Parthasarathy, Face-keg: Fact checking explained using knowledge graphs, in: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021, pp. 526–534.
- [11] A. Saakyan, T. Chakrabarty, S. Muresan, Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic, *arXiv preprint arXiv:2106.03794* (2021).
- [12] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Factify 2: A multimodal fake news and satire news dataset, in: *proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR, 2023.
- [13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.
- [14] S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Findings of factify 2: multimodal fake news detection, in: *proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection*, CEUR, 2023.
- [15] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, et al., Factify: A multi-modal fact verification dataset,

- in: Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY), 2022.
- [16] 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).
  - [17] 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.
  - [18] R. Kumari, A. Ekbal, Amfb: attention based multimodal factorized bilinear pooling for multimodal fake news detection, Expert Systems with Applications 184 (2021) 115412.
  - [19] P. Meel, D. K. Vishwakarma, Han, image captioning, and forensics ensemble multimodal fake news detection, Information Sciences 567 (2021) 23–41.
  - [20] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2560–2569.
  - [21] C. Song, N. Ning, Y. Zhang, B. Wu, A multimodal fake news detection model based on cross-modal attention residual and multichannel convolutional neural networks, Information Processing & Management 58 (2021) 102437.
  - [22] S. R. Sahoo, B. B. Gupta, Multiple features based approach for automatic fake news detection on social networks using deep learning, Applied Soft Computing 100 (2021) 106983.
  - [23] Z. Kang, Y. Cao, Y. Shang, T. Liang, H. Tang, L. Tong, Fake news detection with heterogeneous deep graph convolutional network, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2021, pp. 408–420.
  - [24] L. Hu, T. Yang, L. Zhang, W. Zhong, D. Tang, C. Shi, N. Duan, M. Zhou, Compare to the knowledge: Graph neural fake news detection with external knowledge, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 754–763.
  - [25] Y. Dun, K. Tu, C. Chen, C. Hou, X. Yuan, Kan: Knowledge-aware attention network for fake news detection, in: Proceedings of the AAAI conference on artificial intelligence, volume 35, 2021, pp. 81–89.
  - [26] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, et al., Benchmarking multi-modal entailment for fact verification, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
  - [27] 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).
  - [28] N. Reimers, I. Gurevych, Making monolingual sentence embeddings multilingual using knowledge distillation, arXiv preprint arXiv:2004.09813 (2020).
  - [29] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).
  - [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).