# Countering Dis-information by Multimodal Entailment via Joint Embedding

Aditya Bobde
Dept. of Artificial Intelligence
G H Raisoni College of Engineering
Nagpur, India
bobde_aditya.ai@ghrce.raisoni.net

Pravin Kshirsagar
Dept. of Artificial Intelligence
G H Raisoni College of Engineering
Nagpur, India
pravin.kshirsagar@raisoni.net

Shivam Tawari
Dept. of Artificial Intelligence
G H Raisoni College of Engineering
Nagpur, India
tawari_shivam.ai@ghrce.raisoni.net

Vishal Narnaware
Dept. of Artificial Intelligence
G H Raisoni College of Engineering
Nagpur, India
narnaware_vishal.ai@ghrce.raisoni.net

Adarsh Suryawanshi
Dept. of Artificial Intelligence
G H Raisoni College of Engineering
Nagpur, India
suryawanshi_adarsh.ai@ghrce.raisoni.net

***Abstract*— Abstract Fake news has always been hard to flag and take down before they make a negative impact. With the recent advancement in technology and access to social media platforms, spread of misinformation has picked up dramatically making it near impossible to tag and remove them. Verifying claims is a challenging task and most of the work present is on textual data. However, social media platforms publish multimodal posts often with an image accompanied with a caption. In this work we use Factify dataset for the task of multimodal entailment. We propose a new framework leveraging co-attention layers to jointly understand both the modalities and classify given claims into one of five categories – Support-Multimodal, Support-Text, Insufficient-Multimodal, Insufficient-Text and Refute.**

***Keywords*— *Multimodal Deep Learning, Transformer, Fact Checking, Factify, Co-Attention***

## I. Introduction

Misinformation, false or misleading information, has been a growing concern for society as it makes it hard to find reliable information on the web. Fake news leverages misinformation to claim something which may have never happened or have modified its story often to mudsling a person or group. Recently, a video was circulated online of a person resembling newly appointed United Kingdom's Prime Minister Rishi Sunak where he is seen dancing at a beach club in Ibiza. However, fact checkers [1, 2] rejected the claim and commented that the video dates back to 2019. The person in the video holds a close resemblance to the UK's Prime Minister but is not him. This video went viral even in January 2022 making similar false claims when news started floating for the UK's General Election. Similarly fake news has also been spread in the Healthcare domain, for instance, in 2020 people became skeptical of medical care and even avoided vaccinations as a result of the widespread dissemination of false information concerning COVID-19 [3].

Since ancient times, people have faced false information. With recent advancement in technology and the internet, people have access to a huge amount of information. Social media platforms have made it easier to share news to the masses. They allow users to upload images, videos and textual captions to engage with the users. In general, the information shared on the internet is multimodal. However, this worsens spread of misinformation, making it harder to verify claims in every news shared. Roughly 70 million fake news sites are engaged with Facebook per month [4].

Therefore, various groups have collaborated to protect communities against false information, including journalists, academics, and independent fact-checkers. Many fact checking websites have been created such as FactCheck, Snopes, PolitiFact and Washington Post Fact Checker to help combat fake news. Establishing the disputed claims, obtaining expert perspectives, gathering pertinent data, authenticating sources, looking up any missing data, debating, and finally coming to a conclusion are all typical steps in the fact-checking process. This makes the manual fact checking process tedious and time consuming but accurate. It is impractical for human experts to manually check facts given the volume of content produced daily. Therefore, a lot of academics have been investigating how fact-checking can be automated, using tools like machine learning and natural language processing to predict the accuracy of claims efficiently.

Majority of the work [5, 6, 7, 8, 9] has been done on identifying fake news of textual data. However, most of the data available on the internet is multimodal. In determining whether a piece of information is fabricated or fake, both the modalities – image and text offer insights. The developed model must learn to express content of image and text features as well as their inter-modality interactions. Our focus in this work is multimodal entailment. The objective is to identify multimodal false news, where each data sample includes both a source of accurate information called document and a second source, claim, whose reliability must be confirmed.

## II. Related Work

There are many studies that check the facts and verify the claims made in given text. Liar [5], CREDBANK [6], The Lie Detector [7], MultiFC [8], CheckThat! [9], Claim Matching beyond English[10], FEVER [11], FakeNewsNet [12] and SciFact [13] are some of the datasets that are scrapped from the internet which contain text claims and their document or metadata. Majority of the work is done on verifying textual claims. For instance, a model called ARC-NLP-contra was proposed by team ARC-NLP (Aselsan Research Center - Natural Language Processing) during CLEF 2022 CheckThat! Challenge where they verified tweets by using contradiction check approach [14]. They checked the claims by generating a manually verified facts list from reliable sources. Transformer models have had a huge impact on textual data. Therefore, they have been applied to fact checking tasks as well. LambdaMART which uses ranks predicted by a fine-tuned version of sentence-BERT [40] model, pretrained on Semantic Textual Similarity benchmark (STSb) data, and TF-

IDF was used to check facts [15]. Passive Aggressive Classifier, Bidirectional LSTM [45] and RoBERTa [46] were ensembled to obtain the highest precision for cross lingual fake news detection [16]. Some work has been carried out on comparing Machine Learning models trained with features extracted by classical Doc2Vec algorithm and transformer based architectures such as RoBERTa [46], Electra [47], T5 [48] and Longformers [17]. Another research [18] focuses on using binary and multiclass BERT [37] based text classifiers for identifying articles whose content is irrelevant and for determining truth value respectively. Most of the work leverages deep pretrained language models such as RoBERTa [46], Longformers [17] and T5 [16, 17, 19, 20, 21, 22, 23] for fact checking because of their powerfulness in understanding language and generating texts. However, [24] showed that deep generative models can also be used for the task of claim matching besides encoding-based approaches. To capture long-term relationships between words in phrases, a study [25] that was inspired by Deep Hierarchical Encoder [26] expanded the hierarchical structure of media articles from the article body to the lexical level.

Apart from textual claim verification, there has been limited work done on multimodal fact checking. Factify [27] is one of the multimodal fact checking dataset which contains 50 thousand multimodal claims. This dataset has been used in many studies to verify claims. Most of the studies have extracted features from text and image by using transformer-based architectures such as BERT [37], DeiT [49], DeBERTa [50] and RoBERTa [46], and then concatenated to create a new feature vector, passed through fully connected layers to get final prediction [28, 29, 30]. On the other hand, some studies have focused on incorporating Machine Learning as a final classifier. A decision tree classifier [31] was proposed in [32] which takes text feature as text entailment predicted by BigBird [33], and image feature as similarity score of claim and document image predicted by using ResNet-50 [34]. Another work [35] uses the Resnet-50 [34] model for extracting image features and RoBERTa [46] for getting text features and finally combined to make predictions with Gradient Booster [35]. Another study [36] uses BERT [37] and Vision Transformer [38] to create feature vectors of claim and document text and image respectively. However, they use Conv1d for feature fusion, making the parameters learnable. [39] breaks Factify challenge into text entailment and image entailment task. Sentence BERT [40] and Xception [41] net have been used to generate embeddings. Similarity of these embeddings are calculated using cosine similarity. These features are then fed into two separate fully connected networks to get text and image entailment prediction. Predictions are finally merged in post processing to get final prediction of multimodal entailment. Deep Learning has shown promising impact in healthcare and other domains [42, 43, 44] and recent development in Multimodal data shows great results as well.

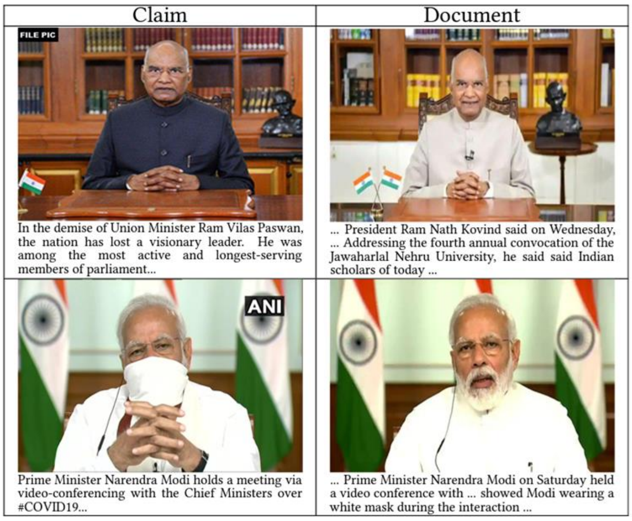
## III. Dataset

Factify [27] is a multimodal dataset of fact checking presented in the De-Factify workshop at AAAI 2022. The dataset contains claims and their respective documents. Each claim has two modalities – text and image. Image text has been extracted by running Google Cloud Vision API's Optical-Character-Recognition system.

The Factify Task is a task that involves detecting fake news. It is modeled as a multimodal entailment, which means that it looks at both the text and the images associated with a reliable source of information (called the "document") and another source whose validity must be assessed (called the "claim"). The goal is to determine if the claim is supported by the document. The task is to predict if a given claim is supported, has no-evidence or refutes the document. Support and no-evidence are further divided into two more categories to take text and image modality into account.

The task is Classifying the data points into one of five categories—Support-Text, Support-Multimodal, Insufficient-Text, Insufficient-Multimodal, and Refute.

The data has been collected from Twitter handles of popular news channels of India and United States: Hindustan-Times and ANI from India, ABC and CNN from US based on accessibility, popularity and tweets per day.



| Claim | Document |
|---|---|
| In the demise of Union Minister Ram Vilas Paswan, the nation has lost a visionary leader. He was among the most active and longest-serving members of parliament... | ... President Ram Nath Kovind said on Wednesday, ... Addressing the fourth annual convocation of the Jawaharlal Nehru University, he said said Indian scholars of today ... |
| Prime Minister Narendra Modi holds a meeting via video-conferencing with the Chief Ministers over #COVID19... | ... Prime Minister Narendra Modi on Saturday held a video conference with ... showed Modi wearing a white mask during the interaction ... |

Fig. 1. Examples of claim and document data sample of Factify Dataset.

Thus, Factify [27] contains five classes and poses a challenge for multimodal entailment. Table 1 contains the description of each class.

TABLE I. Description of Factify Categories

| | | |
|---|---|---|
| **Support-Multimodal** | Text is trusted | Image is trusted |
| **Support-Text** | Text is trusted | Image is neither trusted nor refuted |
| **Insufficient-Multimodal** | Text is neither trusted nor refuted but may have something in common | Image is trusted |
| **Insufficient-Text** | Text is neither trusted nor refuted but may have something in common | Image is neither supported nor refuted |
| **Refute** | Claim text is fake or fabricated | Claim image is fabricated or fake |

## IV. Methodology

### A. Data Efficient Image Transformer

Data Efficient Image Transformer (DeiT) [49] is built upon Vision Transformer (ViT) [38]. DeiT [49] introduces a few modifications on the base architecture, ViT [38], to make it more efficient. The architecture of DeiT [49] is more focused on making it convolution free and to optimize the learning mainly to outperform convolution-based networks.

Transformer-specific teacher-student technique is used to train the model. It is built around a distillation token, which makes sure that the student pays attention and learns from the instructor and does not require a very large amount of data to be trained on.

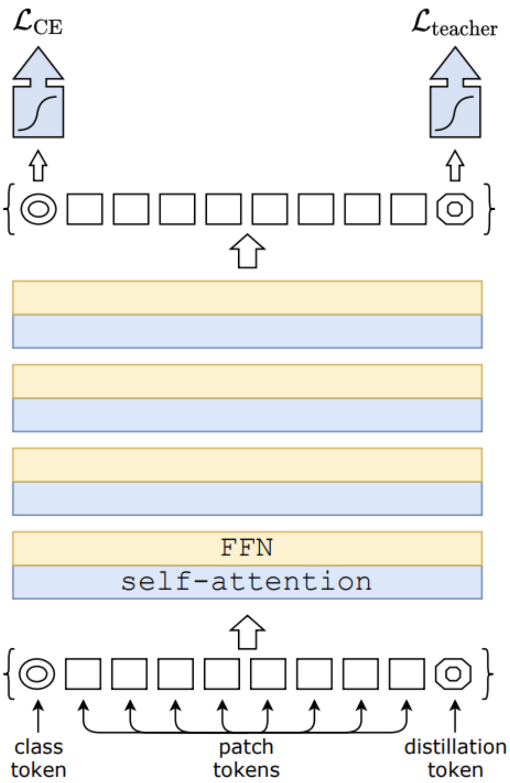


Fig. 2. Architecture of Data Efficient Image Transformer (DeiT)

The image is divided into 16x16 patches like ViT [38]. Then passed through an embedding layer to get fixed size patch embeddings of size d. Class token and Distillation token are added at start and end of the embedding vector respectively along with position embeddings. The architecture of DeiT [49] is shown in figure 2.

DeiT [49] only keeps the output from the Distillation and CLS tokens and discard all the other tokens. Then projects them to the number of classes by running them through two different linear layers. At last, calculates the loss (training), or estimate a class (inference). The model uses Hard Distillation as loss function. In Hard Distillation, the learner tries to imitate the labels that the teacher had expected. In doing so, it lessens the loss of cross-entropy between the labels of the teacher's and the student's softmax. Equation 1 describes the loss function of DeiT [49].

$$\mathcal{L}_{global}^{hardDistill} = \frac{1}{2}\mathcal{L}_{CE}(\psi(Z_s), y) + \frac{1}{2}\mathcal{L}_{CE}(\psi(Z_s), y_t) \quad (1)$$

## B. *Decoding Enhanced BERT with Disentangled Attention*

There are two innovative strategies in Decoding Enhanced BERT with Disentangled Attention (DeBERTa) [50]. There are two vectors that represents encoding of each word and its position respectively, this is called disentangled attention mechanism. Disentangled matrices are used to compute attention weights among words on their contents and relative positions. And second, is an enhanced mask decoder. In order to predict the masked tokens for model pretraining, an improved mask decoder is employed in place of the output softmax layer.

The content and position encoding of each word in DeBERTa [50] is represented by using two vectors, in contrast to BERT [37], which uses an individual vector made up of the word embedding and position embedding of each word in the input layer to represent the word. Disentangled matrices are used to calculate the attention weights between words.

Similar to RoBERTa [46] and BERT [37], DeBERTa [50] is pre-trained using MLM, where a model is trained to anticipate the randomly masked words in a phrase based on the context of adjacent words. DeBERTa [50] sometimes struggles to effectively transfer the context for the masked word prediction since it employs relative locations rather than absolute placements of words. The model must also take into consideration the actual positions in addition to the relative positions. In order to overcome this problem, the absolute positions information is introduced just after the transformer layers and just before the softmax layer, for masked token predictions. They called it as Enhanced Mask Decoding.

## C. *FakeNet*

In this study we started with extraction of feature vectors from the Factify dataset using the Data-Efficient Image Transformer (DeiT) [49] which is trained for image classification tasks. It returned the feature vector of 768 dimensions. For training of the model we had chosen 10% of the data ie 3500 images. For textual data (document and claim) we have used Decoding-enhanced BERT with Disentangled Attention [50] to generate textual data embedding.

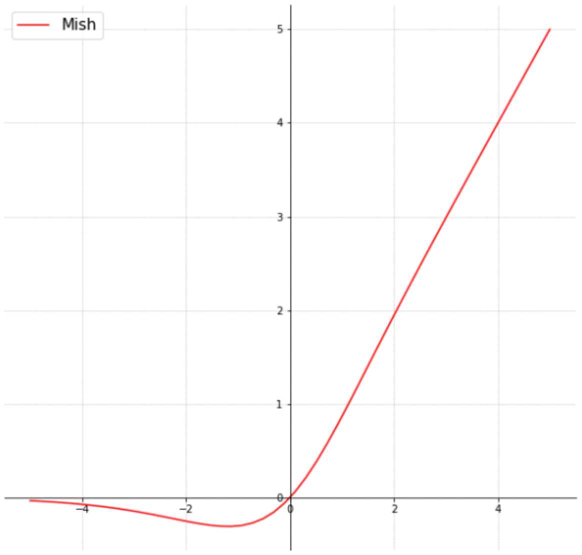


Fig. 3. Mish Activation Function

We demonstrate that these two methods considerably increase the performance of model pretraining and the output in the downstream tasks. We found cosine similarity for both the Claim text and the Document text, which is a measure of similarity between two vectors. Later, it was fed as an input to the model with the feature embedding of the claim text as 768 dimension and the document text embedding as 512

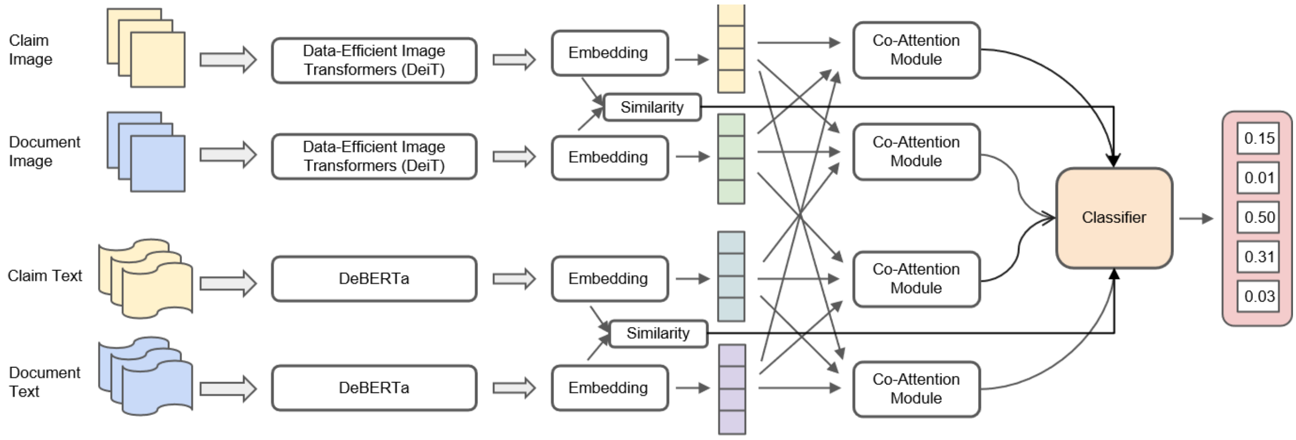


Fig. 4. Architecture of Data Efficient Image Transformer (DeiT)

dimension. The contexts of the text and images are then combined through a number of co-attention modules. Finally, to classify if the data is fake or not, these embeddings are then combined in the final classifier to generate probabilities.

To examine the relationship between a claim and a document, we employ the co-attention layer to fuse 1) pictures of accusations and documents and 2) content of claims and documents independently. Furthermore, the relationship between content and visuals from claims or documents may be interpreted as determining whether or not they are relative. As a result, we use the co-attention layer to combine 3) images and content of claims; 4) images and content of documents. The co-attention technique is used first, and then the aggregation function is used to combine fused tokens into a representative token. That is, we employ mean aggregation to produce R 1xd from a fused embedding with R Nxd, where N is the length of the sequence.

Finally, we combined the picture and text feature embedding into a newly built class called FakeNet. Mish was used as an activation function. Figure 4 shows the graphical plot of Mish activation function. Mish is continuous and differentiable at every point. Mish outperforms both ReLU and Swish, as well as other common activation functions, in various deep networks across hard datasets. The architecture is shown in Fig 4.

## V. Result and Discussion

The results obtained for presented architecture are summarized in table 2. Similarity of claim and document text are calculated by using cosine similarity on text feature vectors which are extracted by DeBERTa [50]. These similarity scores are high with refute class. They are indicating that fake news or claims are generally real news with modified or fabricated information.

TABLE II. METRICS OF MODEL ON VALIDATION SET

| | Precision | Recall | F1 Score |
|---|---|---|---|
| **Support-Multimodal** | 0.56321839 | 0.65333333 | 0.60493827 |
| **Support-Text** | 0.46902655 | 0.35333333 | 0.40304183 |
| **Insufficient-Multimodal** | 0.44886364 | 0.52666667 | 0.48466258 |
| **Insufficient-Text** | 0.49253731 | 0.44 | 0.46478873 |
| **Refute** | 0.96078431 | 0.98 | 0.97029703 |

Furthermore, similarity of other four classes – Support-Multimodal, Support-Text, Insufficient-Multimodal and Insufficient-Text were below 0.50 which is in line with the reasoning that they contain some additional information which cannot be verified with the given document or misses some critical information. This could also be because the claim and document texts have similar words or topics, but not enough to reach the threshold of text entailment.

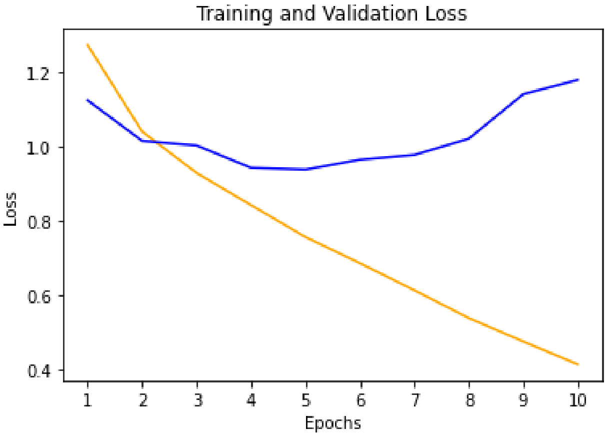


Fig. 5. Training and Validation Loss. Training loss is indicated with orange and validation with blue line.

However, unlike text similarity, there was no particular pattern observed in image similarity. Our architecture has a validation loss of 0.94 and achieves an overall F1 score of 0.60 with refute class F1 score of 0.97. The model can differentiate between fake claims and support or insufficient claims very well but is slightly confused within refined classes.

## VI. Conclusion and Future Work

In this study, the objective was to verify claims using multimodal entailment. The task was to classify a given multimodal claim and document into one of five classes – Support-Multimodal, Support-Text, Insufficient-Multimodal, Insufficient-Text and Refute. We presented a new architecture which takes image and text embeddings along with similarity scores for fact checking. The embeddings for text and image were generated by DeBERTa [50] and DeiT [49] respectively. We found out that using similarity scores along with embeddings improved the score. The task in this paper is far from being finished and requires more attention from research community. Future work could involve experimenting with more multimodal frameworks. Instead of giving a simple yes or no, the level of fakeness could also be determined. Another direction could explain why the claim was predicted in one of the five classes by the model. A possible explanation could solve many of the fact checkers problems.

## References

[1] S. Das, “Old Video Shared As Rishi Sunak Offering Prayers To Cow After Becoming PM | BOOM,” www.boomlive.in, Oct. 29, 2022. https://www.boomlive.in/fact-check/viral-video-united-kingdom-prime-minister-rishi-sunak-cow-worship-social-media-19748 (accessed Nov. 29, 2022).

[2] J. Hassan, “Old Video Of Rishi Sunak Worshiping Cow Viral With Misleading Claim!orDid This Viral Video Show Rishi Sunak Worshiping Cow After Becoming PM? No, Viral Claim Is Misleading,” thelogicalindian.com, Oct. 31, 2022. https://thelogicalindian.com/fact-check/old-video-of-rishi-sunak-worshiping-cow-viral-with-misleading-claim-38378 (accessed Nov. 29, 2022).

[3] Md Saiful Islam, Tonmoy Sarkar, Sazzad Hossain Khan, Abu-Hena Mostofa Kamal, SM Murshid Hasan, Alamgir Kabir, Dalia Yeasmin, Mohammad Ariful Islam, Kamal Ibne Amin Chowdhury, Kazi Selim Anwar, et al. 2020. Covid-19–related infodemic and its impact on public health: A global social media analysis. The American journal of tropical medicine and hygiene, 103(4):162

[4] X. Zeng, A. S. Abumansour, A. Zubiaga, Automated fact-checking: A survey, 2021. arXiv:2109.11427.

[5] W. Y. Wang, ” liar, liar pants on fire”: A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).

[6] T. Mitra, E. Gilbert, Credbank: A large-scale social media corpus with associated credibility annotations, in: ICWSM, 2015.

[7] R. Mihalcea, C. Strapparava, The lie detector: Explorations in the automatic recognition of deceptive language, in: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort ’09, Association for Computational Linguistics, USA, 2009, p. 309–312.

[8] I. Augenstein, C. Lioma, D. Wang, L. Chaves Lima, C. Hansen, C. Hansen, J. Grue Simonsen, Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, in: EMNLP, Association for Computational Linguistics, 2019.

[9] Barrón-Cedeño, A., Elsayed, T., Nakov, P., Da San Martino, G., Hasanain, M., Suwaileh, R., Haouari, F., Babulkov, N., Hamdan, B., Nikolov, A. and Shaar, S., 2020, September. Overview of CheckThat! 2020: Automatic identification and verification of claims in social media. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 215-236). Springer, Cham.

[10] A. Kazemi, K. Garimella, D. Gaffney, S. A. Hale, Claim matching beyond english to scale global fact-checking, 2021. arXiv:2106.00853.

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

[12] 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, 2019. arXiv:1809.01286.

[13] Wadden, D., Lin, S., Lo, K., Wang, L.L., van Zuylen, M., Cohan, A. and Hajishirzi, H., 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

[14] Toraman, C., Ozcelik, O., Şahinuç, F. and Sahin, U., 2022. ARC-NLP at CheckThat! 2022: contradiction for harmful tweet detection. Working Notes of CLEF.

[15] Chernyavskiy, A., Ilvovsky, D. and Nakov, P., 2021, January. Aschern at CLEF CheckThat! 2021: lambda-calculus of fact-checked claims. In CLEF (Working Notes).

[16] Arif, M., Tonja, A.L., Ameer, I., Kolesnikova, O., Gelbukh, A., Sidorov, G. and Meque, A.G., 2022. CIC at CheckThat! 2022: multi-class and cross-lingual fake news detection. Working Notes of CLEF.

[17] CIVIC-UPM at CheckThat! 2021: Integration of Transformers in Misinformation Detection and Topic Classification

[18] Blanc, O., Pritzkau, A., Schade, U. and Geierhos, M., 2022. CODE at CheckThat! 2022: multi-class fake news detection of news articles with BERT. Working Notes of CLEF.

[19] Pritzkau, A., Blanc, O., Geierhos, M. and Schade, U., 2022. NLytics at CheckThat! 2022: hierarchical multi-class fake news detection of news articles exploiting the topic structure. Working Notes of CLEF.

[20] Du, S.M., Gollapalli, S.D. and Ng, S.K., 2022. NUS-IDS at CheckThat! 2022: identifying check-worthiness of tweets using CheckthaT5. Working Notes of CLEF.

[21] Zhuang, Y. and Zhang, Y., 2022. Yet at Factify 2022: Unimodal and bimodal roberta-based models for fact checking. In Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR.

[22] Kshirsagar, P.R., Akojwar, S.G. and Bajaj, N.D., 2018. A hybridised neural network and optimisation algorithms for prediction and classification of neurological disorders. *International Journal of Biomedical Engineering and Technology*, *28*(4), pp.307-321.

[23] Hasanin, T., Kshirsagar, P.R., Manoharan, H., Sengar, S.S., Selvarajan, S. and Satapathy, S.C., 2022. Exploration of Despair Eccentricities Based on Scale Metrics with Feature Sampling Using a Deep Learning Algorithm. *Diagnostics*, *12*(11), p.2844.

[24] RIET Lab at CheckThat! 2022: Improving Decoder based Re-ranking for Claim Matching

[25] Kumar, S., Kumar, G. and Singh, S.R., 2022. TextMinor at CheckThat! 2022: fake news article detection using RoBERT. Working Notes of CLEF.

[26] S. Yoon, K. Park, J. Shin, H. Lim, S. Won, M. Cha, K. Jung, Detecting incongruity between news headline and body text via a deep hierarchical encoder, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 791–800.

[27] 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: De-Factify workshop at AAAI, 2022.

[28] Hulke, N., Siva, B.R., Raj, A. and Saifee, A.A., 2021. Tyche at Factify 2022: Fusion Networks for Multi-Modal Fact-Checking.

[29] Wang, W.Y. and Peng, W.C., 2022. Team Yao at Factify 2022: Utilizing Pre-trained Models and Co-attention Networks for Multi-Modal Fact Verification. arXiv preprint arXiv:2201.11664.

[30] Zhuang, Y. and Zhang, Y., 2022. Yet at Factify 2022: Unimodal and bimodal roberta-based models for fact checking. In Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830

[32] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, A. Bandhakavi, Logically at the factify 2022: Multimodal fact verfication, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022

[33] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences, Advances in Neural Information Processing Systems 33 (2020) 17283–17297.

[34] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[35] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, Advances in neural information processing systems 30 (2017) 3146–3154

[36] Bai, W., 2022. Greeny at Factify 2022: Ensemble model with optimized roberta for multi-modal fact verification. In Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR.

[37] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, ICLR (2021)

[39] A. Dhankar, O. Zaiane, F. Bolduc, UofA-Truth at Factify 2022 : A simple approach to multi-modal fact-checking, in: Proceedings of De-

Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022

[40] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bertnetworks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[41] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258

[42] Manoharan, H., Rambola, R.K., Kshirsagar, P.R., Chakrabarti, P., Alqahtani, J., Naveed, Q.N., Islam, S. and Mekuriyaw, W.D., 2022. Aerial Separation and Receiver Arrangements on Identifying Lung Syndromes Using the Artificial Neural Network. *Computational Intelligence and Neuroscience*, *2022*.

[43] Kshirsagar, P.R., Akojwar, S.G. and Dhanoriya, R.A.M.K.U.M.A.R., 2017. Classification of ECG-signals using artificial neural networks. In *Proceedings of International Conference on Intelligent Technologies and Engineering Systems, Lecture Notes in Electrical Engineering* (Vol. 345).

[44] Kollu, P.K., Kumar, K., Kshirsagar, P.R., Islam, S., Naveed, Q.N., Hussain, M.R. and Sundramurthy, V.P., 2022. Development of advanced artificial intelligence and IoT automation in the crisis of COVID-19 Detection. *Journal of Healthcare Engineering*, *2022*.

[45] Chiu, J.P. and Nichols, E., 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the association for computational linguistics*, *4*, pp.357-370.

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

[47] Clark, K., Luong, M.T., Le, Q.V. and Manning, C.D., 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

[48] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, *21*(140), pp.1-67.

[49] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H., 2021, July. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning* (pp. 10347-10357). PMLR.

[50] He, P., Liu, X., Gao, J. and Chen, W., 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.