Logically at Factify 2: A Multi-Modal Fact Checking System Based on Evidence Retrieval techniques and Transformer Encoder Architecture

Pim Jordi Verschuuren, Jie Gao, Adelize van Eeden, Stylianos Oikonomou and Anil Bandhakavi

Brookfoot Mills, Brookfoot Industrial Estate, Brighouse, HD6 2RW, United Kingdom

Abstract

In this paper, we present the Logically submissions to De-Factify 2 challenge (DE-FACTIFY 2023) on task 1 of Multi-Modal Fact Checking. We describe our submission to this challenge including explored evidence retrieval and selection techniques, pre-trained cross-modal and unimodal models, and a cross-modal veracity model based on the well established Transformer Encoder (TE) architecture which heavily relies on the concept of self-attention. Exploratory analysis is also conducted on the Factify 2 data set that uncovers the salient multi-modal patterns and hypothesis motivating the architecture proposed in this work. A series of preliminary experiments were done to investigate and benchmark different pre-trained embedding models, evidence retrieval settings and thresholds. The final system, a standard two-stage evidence based veracity detection system, yielded a weighted average F1 score of 0.79 on both the validation set and final blind test set of task 1, which achieved 3rd place with a small margin to the top performing systems on the leaderboard among 9 participants.

Keywords

fact verification, multimodal representation learning, multimodal entailment, text entailment, Multi-head Attention

1. Introduction

Misinformation and fake news can spread rapidly and cause harm at various levels. One way to protect ourselves from these negative impacts is through fact-checking and debunking false information with evidence-based reporting. However, this process can be resource-intensive and time-consuming. To address this issue, researchers have developed automated fact-checking systems using deep learning techniques, which can handle tasks such as claim detection, claim matching, evidence retrieval, and veracity prediction using natural language processing techniques on textual content. While there has been progress in this area, there is still a need for multimodal approaches that can handle both text and image inputs. To address this gap, this paper presents a multimodal veracity prediction system for automated fact-checking which

was developed as part of the Factify 2 competition organized by De-Factify@AAAI 2023.

The remainder of the paper is structured as follows: Section 2 presents a brief overview of related work and section 3 describes our general framework and model architecture. Section 4 discusses the dataset supplied by the Factify 2 competition followed by an overview of our experiments in section 5. Section 6 and 7 present the final results and our conclusions, respectively.

2. Related Work

As an essential part of automated fact verification, effective techniques for modeling claim-evidence for veracity prediction have been a hot topic and key research questions in existing fact-checking methods. Most of the recent work focuses on using textual evidence in veracity prediction of which there are mainly two lines of work. One direction [1, 2, 3] is to use a single document (such as is provided in the Factify task dataset) with long text evidence and through leveraging models constructed for long sequences. Examples such as BigBird [4], Longformer[5] and recent advancements in the ConvNets architecture witnessed in the Long Range Arena leaderboard (e.g., Mega [6], S5[7]) are seen to obtain top results in a wide range of tasks and other leaderboards. The benefits of exploiting long-sequence models at document level is a) the simplicity of the overall architecture; b) allows to accommodate for more context of the whole article into modeling and natural language inference. An optimal setup of the maximum length for both claim (or query) and document sequence, and the document level veracity labels is commonly required [8, 1, 3]. The advantage of incorporating lots of context into inference is also seen in modeling question answering (QA) tasks [4, 5], for which the document-level veracity labels are relatively "cheap" to obtain. The downside of using a simple long-text model technique at document-level is the lack of interpretability (w.r.t. evidence selection), it is computational expensive, the limitation in dealing with the complexity of certain (multi-hop) claims [9], and lack of diversity and scalability when dealing with a large amount of diverse documents in a real-world application. These constraints were more apparent in open domain fact checking tasks that make use of web data extracted with commercial search engines as building blocks in fact-checking systems in order to incorporate more diverse sources. It is worth to note that long-sequence models can be adapted for the purpose of evidence selection e.g., through framing the task as a token-level prediction task. For instance, as one of the top systems in the SciFact leaderboard [1], LongChecker [10] used LongFormer [5] for scientific claim verification with paragraph-level evidence selection. In their method, every sentences is inserted with a [CLS] token with global attention, which allows the model to predict on this sentence-level token as evidence. Most of these works focus on a limited context such as a few Wikipedia documents, a single article and abstracts or text snippets from either research literature or a small synthetic corpus.

Another line of work widely adopted and one of the key tasks in FEVER [11, 12] is to involve evidence retrieval and selection. The framework exploits larger document context to extract evidentiary (or rationales) passages as first step and veracity prediction is then modeled to condition on the claim and the selected rationales. The evidentiary passages report the

---

[1]https://leaderboard.allenai.org/scifact/submissions/public

findings to the claim which can be used to justify each veracity label and can be selected at either sentence- or paragraph-level. Despite the revolutionary breakthroughs with Large-Scale Language Models (LSLMs), such as GPT-3[13] and ChatGPT[2], and their impressive generative capabilities, these large models are still lacking key zero-shot or few-shot learning capabilities needed for fact checking tasks. This is mainly due to their incorrectly retrieved, incomplete or outdated knowledge stored in their weights which makes these techniques susceptible to hallucinations [14, 15], which is conflicting with fact checking tasks that require factuality as an essential element in modeling. Moreover, an efficient approach to keep LSLMs up-to-date and grounded to ever-growing factual and new information is imperative but still unresolved to date. Recent work [15, 16] shows that lightweight methods with fine-tuned and smaller models outperform these big models in a range of knowledge-intensive NLP tasks including Natural Language Inference (NLI), Recognizing textual entailment (RTE), Reading Comprehension (RC), QA, etc. Sentence-BERT (SBERT) [17] is one of the most popular techniques based on the BERT language model [18] used for evidence selection [19, 20] which can be framed as a sentence-pair regression task. SBERT models are used to encode contextualized representations for each of the evidence passages which are then ranked according to their semantic similarity with the contextualized representation of the corresponding claim. In the final step, top k evidentiary passages are selected for veracity prediction. The challenge of this multi-staged verification framework is 1) the rationales extracted out-of-context may lack information required to make a prediction (e.g., acronyms, unresolved coreferences); 2) the evidence extraction (through passages ranking) requires high quality training data that is costly to obtain with domain experts from both closed and open domain tasks [21]. Various efforts to address the constraints have been undertaken to explore 1) paragraph level train data from scientific literature with paper title as claim and abstract as evidence as high-precision heuristics (e.g., SciFact [1]); 2) QA dataset with question and answer considered as claim and evidence respectively [22]; 3) NLI dataset with the claim as hypothesis and evidence as evidence [23]. We follow a second line of work for which the evidence retrieval component is implemented in our system following current SoTA methods.

Automated multi- or cross-modal fact checking is an underdeveloped field compared to text-based techniques. Recent developments have shown that cross-modal pre-trained models (e.g.,VideoBERT [24], VisualBERT [25], Uniter [26], CLIP [27]) have achieved significant results in downstream cross-modal tasks [28, 29, 30] with great transferability for zero-shot or few-shot scenarios. Our work is inspired by [31], which was one of the initial explorations in multimodal fact-checking task. In their proposed method, the Contrastive Language−Image Pre-training (CLIP) model [27]) is adopted as encoder to learn joint language-image embedding between each image and input claim text. Top-5 candidate image evidences are taken as input along with multimodal claim for multimodal claim verification model with a simple cross-attention network. It is worth noting that the CLIP model allows to model image-text contextual alignment at coarse-grained contextual (global) level but ignores the compositional matching of disentangled concepts (i.e., finer-grained cross-modal alignment at region-word level)[30, 32, 32].

---

[2]https://openai.com/blog/chatgpt/

3. Methodology

3.1. Problem statement

We frame the Factify 2 problem as a multimodal entailment task as in the previous submission [3], which considers a multimodal claim $c = c_{text} + c_{image}$ as hypothesis and a multimodal document $d = d_{text} + d_{image}$ as premise. The goal is to learn a function $f(c, d)$ that infers one of the five entailment categories including "Support_Multimodal", "Support_Text", "Refutes", "Insufficient_Multimodal" and "Insufficient_Text". Additional details on the task can be found in [33].

3.2. General Architecture

Our system architecture follows a standard two-stage claim verification approach as established through various shared tasks in recent years, typically FEVER[34], FEVER 2.0 [35], FEVEROUS [36] and SCIVER [37]. First, a textual evidence retrieval component identifies from a given document the evidence passages most relevant to the corresponding claim text. Then, a transformer based cross-modal model is trained on all the input across multimodalities including selected evidence passages text, claim text, claim image, document image, claim OCR text and document OCR text to predict five multimodal entailment categories with respect to the multimodal claim. A pre-trained cross-modal model (i.e. CLIP) and a pre-trained text embedding model are both employed in the embedding layer in order to learn a cross-modal matching model using both unified-multimodal and unimodal representations. Overall, the implemented architecture adopts a list-wise concatenation strategy [38] which is one of common strategies in most recent sequence-to-sequence SoTA veracity prediction models.
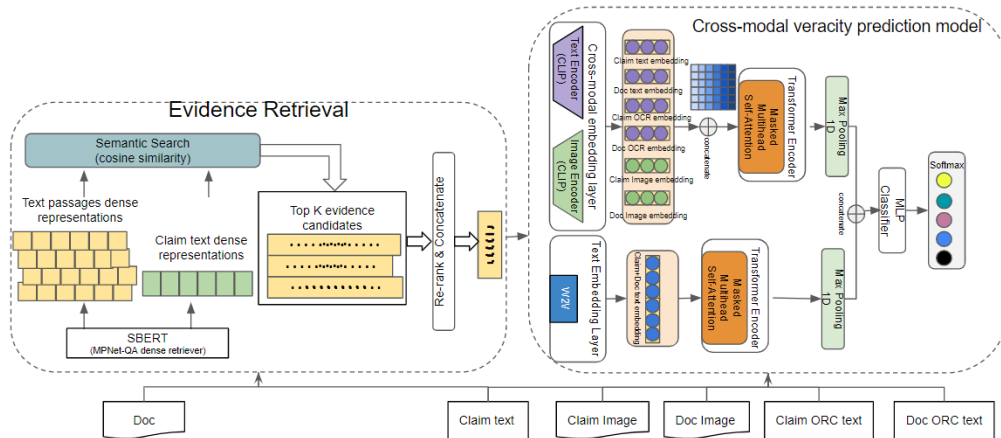

Figure 1: Logically General System Architecture

3.3. Evidence Retrieval

In evidence retrieval, 'multi-qa-mpnet-base-dot-v1' [3] and is employed to compute embeddings for both claim text and document text at passage level. In terms of passage granularity, both paragraph- and sentence-level retrieval have been experimented with (see Section 5). This is a SBERT model based on the MPNet architecture [39] and is trained on a Question-Answer (QA) dataset with 215M QA pairs from diverse sources. The model was tuned for a semantic search using a dot-product score function in order to find relevant passages corresponding to a given query. The model encodes text into a 768-d vector and supports 512 maximum number of tokens.

Regarding the similarity computation and semantic search, we use a simple dot product with the normalised SBERT embeddings (as proxy to cosine similarity) which enables a quick and efficient passage ranking and scalability of up to about 1 Million entries.

Top K passages obtained from the semantic search are then re-ranked based on their relevancy to the claim text and concatenated into a longer text snippet before being fed into the cross-modal veracity prediction model.

3.4. Embedding Layer

Our embedding layer consists of a cross-modal encoder and a unimodal text encoder. We hypothesize that modeling solely on text-to-text interaction (i.e., text premise and hypothesis) can supplement the modeling solely on cross-modal premise and hypothesis interaction and vice versa. This architecture facilitates the measuring of multimodal semantic relatedness in this multimodal fact checking task by mapping more textual alignment signals into subsequent semantic space. This considers that text specific models can capture more accurate and semantically meaningful word- or sentence-level alignment.

The cross-modal encoder is implemented with a pre-trained CLIP model that aims to map visual and text embeddings into a common space. The ViT-B/32 variant (ViT-Base with patch size 32) is chosen in this work because of its smaller amount of parameters, less FLOPS and greater inference speed. ViT-B/32 consists of a text encoder and an image encoder which are used to encode text inputs (including claim text, evidentiary passage and two images OCR text) and image inputs (including claim image and document image) respectively before concatenating into a 6×512 matrix as a single input to the subsequent transformer encoder. The CLIP architecture allows for a maximum input text length of 77 tokens. The pre-trained Word2vec model ("Word2vec Google News 300") [40] is adopted as a unimodal text encoder. It encodes the concatenated text sequence of claim and document evidentiary passage text, and obtains a 300-D feature vector for each token. Zero-padding is applied to match the longest sentence in the training set. Both the pre-trained CLIP and Word2Vec embedding model were not fine-tuned.

3.5. Cross-modal veracity prediction

The second component of veracity prediction is based on the well established Transformer Encoder (TE) architecture, which heavily relies on the concept of self-attention [41] to effectively

---

[3]The model is available on the Hugging Face hub, accessible via https://huggingface.co/sentence-transformers/multi-qa-mpnet-base-dot-v1

model higher-order interactions and context in an input. Recent research has shown that multi-head self-attention mechanisms and transformer architectures are computationally efficient and accurate in this regard. The self-attention mechanisms of the TE encoder allows for simple but powerful reasoning that can identify hidden relationships between vector entities, regardless of whether they are visual or textual in nature. Therefore, our cross-modal veracity prediction model is implemented based on self-attention mechanisms to learn the joint distribution of text representations of claim-document text pair and cross-modal feature representations of all modalities contained in claim and document.

Specifically, the claim and document embeddings of joint input by CLIP and text input by text embedding layer are passed through two separate transformer encoders [41] consisting of N identical sequential blocks of a multi-head attention (MHA) and a fully connected feed-forward network (FFN). Within each transformer encoder, multiple blocks allows for a deeper understanding of the inputs. For each block the input x is passed through a multi-head attention layer of which the output is added to the initial input such that information in the initial sequence is not lost. Layer normalization is applied to the output to allow for faster training and small regularization i.e. $x = \text{LayerNorm}(x + \text{MHA}(x))$. The output is then passed to a feed-forward network to allow for more model complexity. The output is again added to the original input and layer normalization is applied i.e. $x = \text{LayerNorm}(x + \text{FFN}(x))$. The output of the final block (i.e., the output of each transformer encoder in the diagram) is passed through an adaptive max pooling layer to reduce the output dimensions. The output of two separate transformer encoders are then concatenated before feeding into a MLP classifier for the five category prediction. The five categories probabilities are obtained from the final output softmax layer.

4. Factify Dataset

4.1. Dataset Description

The Factify 2 dataset created and supplied by the organisers covers a train, validation, and test set. The train set contains 35000 data pairs, while the validation and test sets each contain 7500 data pairs. Each data pair consists of a claim and a document, each of which comprises an image, a text, and an OCR text extracted from the image. The data pairs are annotated with one label from 5 categories including Support_Multimodal, Support_Text, Refute, Insufficient_Multimodal, or Insufficient_Text.

4.2. Text Length Distribution

The training set text and OCR text length distributions are represented in Figures 2 and 3. The text length distribution varies between the claim and document text, with the document text that tends to be much longer. This is expected as it is used to verify the claim. From Figure 2 (a), we can can see that claim text is much shorter and less varied for the Refute category than for the rest of the categories, which all have similar claim text length distributions. Figure 2 (b) shows that the Support_Multimodal and Support_Text categories have the larger spread of document text lengths and also the longest document text lengths. The two Insufficient

categories have on average a smaller document text length, and Refute has the smallest variance and maximum length in document text length.

Considering the claim OCR length we see from Figure 3 that the Refute category has a much larger claim OCR length distribution and maximum length than any other category. The second largest claim OCR length distributions are the Support_Text and the Insufficient_Text categories, which then leaves the two Multimodal categories with the shortest claim OCR text lengths. The document OCR length distribution is very similar to that of the claim OCR, from Figure 3b we see the only real difference is that the two Text categories have a smaller document OCR length distrubution than that of the claim OCR.
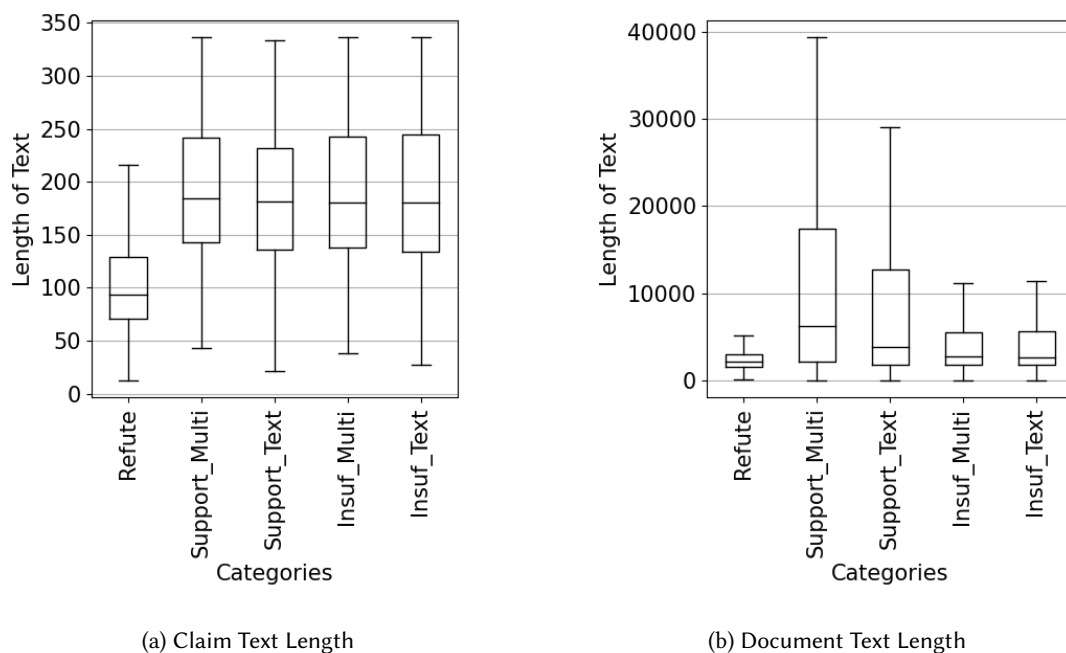

(a) Claim Text Length

(b) Document Text Length

Figure 2: Boxplot of Text Length Distribution of all Categories

4.3. Image Similarity Distribution

An image similarity investigation was conducted in order to gain an intuition of the similarity between the claim and document images for each category. Using image pairwise CLIP embeddings we calculate a similarity score and analyse it per category. Figures 4a and 4b illustrate that the similarity between the claim and document image is comparatively higher within the categories for Support_Multimodal and Insufficient_Multimodal than the other categories. The label correlation with similarity of image pairs has largely increased compared to factity 1 dataset [3] of last year. This further indicates that there is explicit correlation within the multimodal categories which can be leveraged to learn and verify multimodal entailment categories.
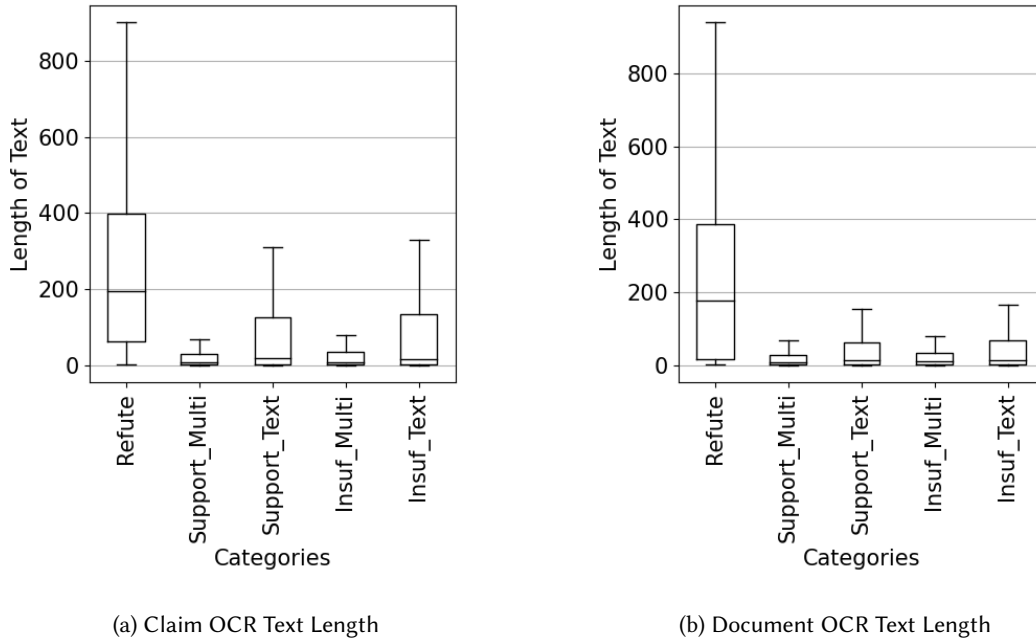
(a) Claim OCR Text Length   (b) Document OCR Text Length

Figure 3: Boxplot of OCR Text Length Distribution of all Categories

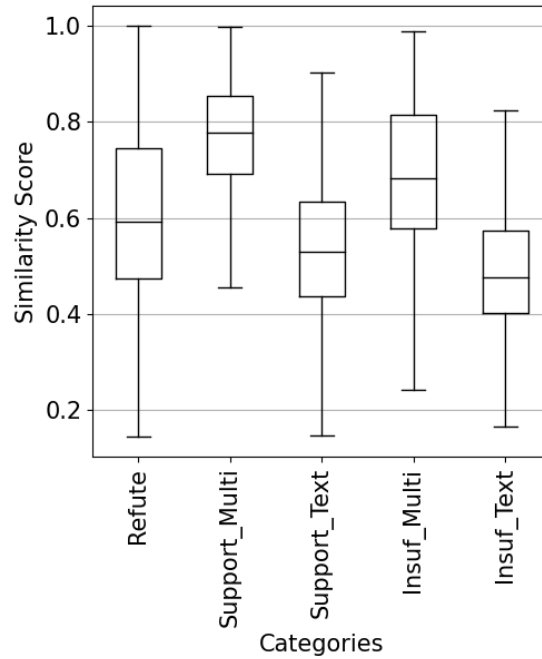## 4.4. Multimodal Similarity Distribution

The multimodal CLIP similarity among multimodal claim and doc pairs is explored to investigate our hypothesis that doc image should contain content that is related to the claim in order to entail either support or refute verdict decisions. Figures 5a and 5b depict the cosine similarity scores between the claim text and document image. It is noticeable that "Support_Multimodal" presents the highest pairwise similarity correlation between label and claim-evidence pair. "Insufficient text" have the lowest pairwise similarity correlation, although our initial hypothesis was that "Insufficient_Multimodal" should have the lowest value. This analysis suggests that differentiating between the different categories based on the claim text and document image correlation could be challenging.

In terms of correlation between the claim image and document text, due to the maximum text sequence constraints with CLIP, text access maximum length is truncated. Consequently, longer context of document text is not incorporated in this analysis. As shown in Figure 6a and 6c, there is low degree of similarity correlation across the five categories, among which the "Refute" category shows highest similarity correlation.

Lastly, Figure 6b and Figure 6d show the similarity correlation between the claim image and the claim text, and show no significant deviation in similarity scores of different categories when the claim image and claim text are compared to each other. For the purpose of this task and this dataset, we hypothesize that the claim image should provide supplementary information to the claim text.
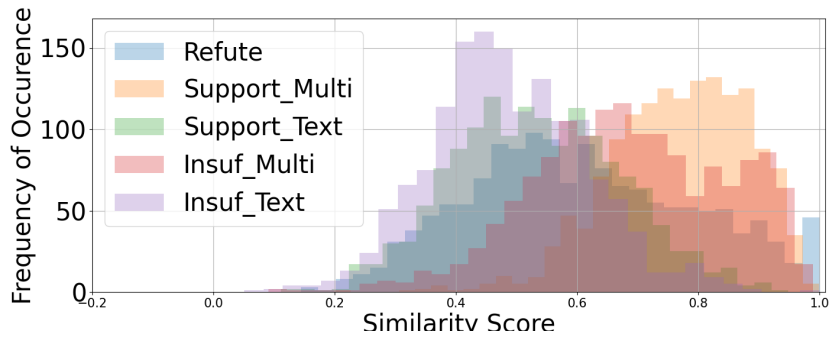
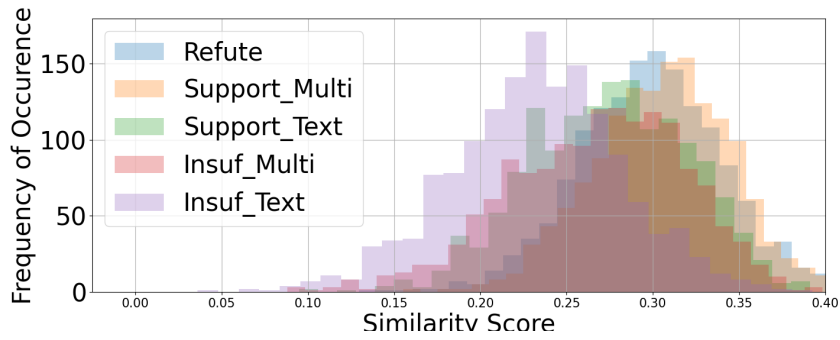(a) Claim Image and Document Image Similarity Score Histogram

(b) Claim Image and Document Image Similarity Boxplot

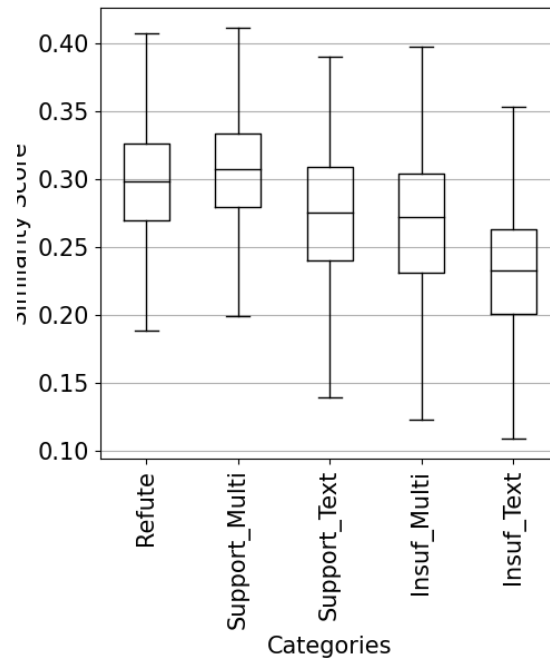**Figure 4:** Claim Image and Document Image Similarity Distribution

5. Experiments

5.1. Model settings

To validate and optimize the effect of evidence retrieval, we attempt to experiment with our model with 1) including or excluding evidence selection; 2) varying the length of evidence doc text sorted by evidence retriever; 3) passage ranking at paragraph level versus sentence level; 4) text-to-text alignment with SBERT versus cross-modal alignment with CLIP. Both SBERT and CLIP is used to rank evidence doc with paragraph and sentence level; 5) if SBERT model trained on QA dataset perform better than general purpose SBERT model. Note that ranking

(a) Claim Text and Document Image Similarity Score Histogram

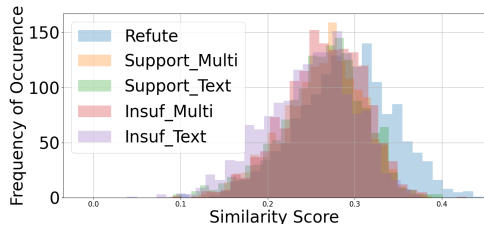(b) Claim Text and Document Image Similarity Boxplot

Figure 5: Claim Text and Document Image Similarity Scores

at paragraph level on top <5 or sentence level on top <5 is only an option for CLIP due to its maximum allowed length restriction.

For two transformer encoders, we choose an empirical setting of four heads in two MHAs. The number of sequential MHA and feed-forward network blocks per embedding input is $N_{blocks} = 2$. All our experiments are trained on a 3-layered MLP and the number of nodes per layer are set to 3072, 1024 and 5, respectively. A dropout of 0.5 and ReLU activations are applied between the MLP layers.

Preliminary experiments conducted in this work are elaborated in details as follows:
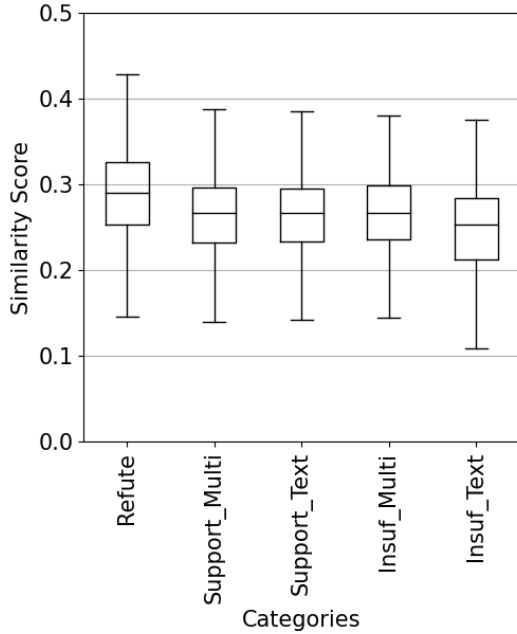
- "model_w/o_ER": to validate the effectiveness with evidence retrieval, we remove evidence
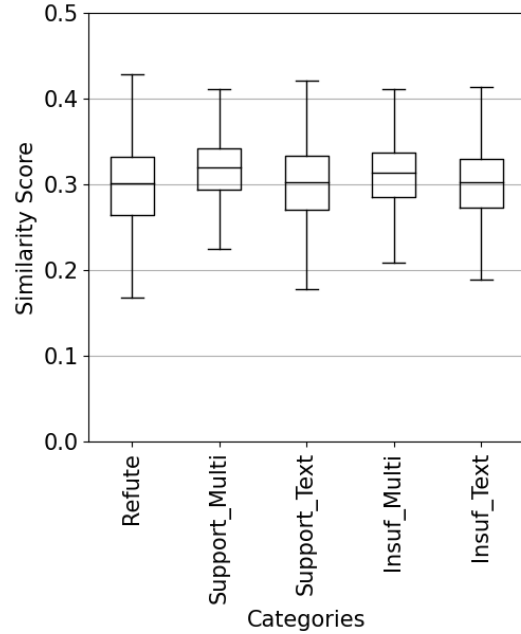
(a) Claim Image and Document Text Similarity Score Histogram
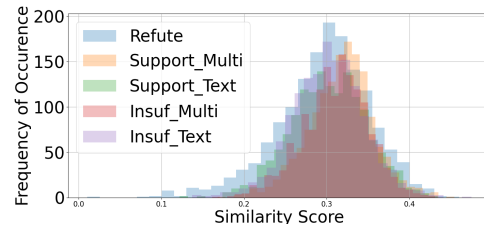
(b) Claim Image and Claim Text Similarity Score Histogram

(c) Claim Image and Document Text Similarity Score Boxplot

(d) Claim Image and Claim Text Similarity Score Boxplot

Figure 6: Image and Text Similarity distribution among multimodal claim and doc

retrieval in our system and provide original document text to "Cross-modal veracity prediction model".

- "SBERT_sentence_ER_top5": One of the "top" [4] performing general purpose SBERT model ("all-MiniLM-L6-v2") is chosen in our experiment. This is an all-round model tuned for many use cases and 5 times faster while offering good quality compared to the best all-round model "all-mpnet-base-v2". The model is trained on a large and diverse dataset of over 1 billion training pairs and also fine-tuned for dot-product score function suitable for cosine similarity. The use of the all-round model allows us to evaluate the value of adopting QA fine-tuned counterpart that we hypothesize to be the optimal solution. Top

---

[4]The best performing general purpose model is selected with a sorted list of model performances and recommended use cases provided by SBERT, accessible via https://www.sbert.net/docs/pretrained_models.html

5 sorted sentences sorted by all-round SBERT model is configured in this setting.

- "SBERT_sentence_ER_top10": Top 10 sorted sentence sorted by all-round SBERT model is configured in this setting.
- "SBERT_sentence_ER_top15": Top 15 sorted sentence sorted by all-round SBERT model is configured in this setting.
- "SBERT-QA_paragraph_ER_top5": SBERT QA dataset fine-tuned model (as described in 3.3) is adopted in this setting to obtain top 5 paragraphs as evidentiary passages for veracity inference in this setting.
- "SBERT-QA_sentence_ER_top5": Top 5 sentences sorted by SBERT QA model and selected as evidentiary passages in this setting.
- "BigBird_w/o_ER": To evaluate the value of evidence selection against the long context modeling solution, the Google's BigBird pre-trained model fine-tuned on Factity dataset from last year [3] is used to replace the Word2Vec model in the "Text Embedding layer" with this setting. This BigBird model allows a maximum 1396 tokens and contextual representation of text is adopted in this setting.

5.2. Training and validation

For our experiment, the model was trained up to 80 epochs with early stopping on minimum validation loss by minimizing the cross-entropy loss function using the adaptive AdamW optimizer [42] with an initial learning rate of $\gamma = 1e-4$ and epsilon $\epsilon = 1e-8$ with batch size $N_{batch} = 16$. Early stopping patience is set to 5. A linear decreasing learning rate scheduler was used including $N_{steps} = 438$ warming up training steps during which the learning rate increased linearly to the chosen learning rate.

We have found that data scraping errors lead to invalid doc text content in the development dateset provided by organiser with 463 and 114 invalid samples in train and val set respectively. There also are 112 invalid samples in test set. This results in document text containing only "We've detected that JavaScript is disabled in this browser ...". The invalid samples are removed from our training data.

6. Results and Discussion

The best model results in preliminary experiments described in section 5 are presented in Table 1, Table 2 and Table 3 respectively.

Firstly, the Table 1 shows that our veracity model without ER exhibits a reasonably good performance and utilising the long sequence model (BigBird) for text embeddings improves the base model with a small margin, by 1% for all categories except "Refute". As comparison, further experiments with ER are conducted of which the results are presented in Table 2 and Table 3. The results in Table 2 indicate that all-round SBERT based evidence selection does not provide obvious performance improvement based on current preliminary explorations covering three top K sentences settings (K=5, 10, 15). In contrast, SERT-QA based model achieves big marginal improvement at both paragraph and sentence level. Our experiments covers both top 5 paragraphs and sentences, which improves best base model (without ER) by

Table 1

5-way Classification Results of experiments without ER on val set

Categories	model_w/o_ER			BigBird_w/o_ER		
	P	R	F1	P	R	F1
Support_Multimodal	0.73	0.79	0.76	0.73	0.81	**0.77**
Support_Text	0.71	0.61	0.66	0.77	0.59	**0.67**
Insufficient_Multimodal	0.66	0.66	0.66	0.64	0.70	**0.67**
Insufficient_Text	0.71	0.75	0.73	0.73	0.75	**0.74**
Refute	0.99	0.98	0.98	0.98	0.98	0.98
Weighted Avg.	0.76	0.76	0.76	0.77	0.77	**0.77**

Table 2

5-way Classification Results of experiments with all-round SBERT + ER on val set

Categories	SBERT_sentence_ER_top5			SBERT_sentence_ER_top10			SBERT_sentence_ER_top15		
	P	R	F1	P	R	F1	P	R	F1
Support_Multimodal	0.72	0.85	0.78	0.74	0.78	0.76	0.75	0.77	0.76
Support_Text	0.63	0.73	0.68	0.71	0.61	0.66	0.71	0.62	0.66
Insufficient_Multimodal	0.70	0.64	0.67	0.66	0.67	0.66	0.65	0.67	0.66
Insufficient_Text	0.80	0.58	0.67	0.70	0.77	0.74	0.71	0.76	0.73
Refute	0.96	0.99	0.97	0.96	0.99	0.97	0.98	0.98	0.98
Weighted Avg.	0.76	0.76	0.75	0.76	0.76	0.76	0.76	0.76	0.76

Table 3

5-way Classification Results of experiments with SBERT-QA + ER on val set

Categories	SBERT-QA_paragraph_ER_top5			SBERT-QA_sentence_ER_top5		
	P	R	F1	P	R	F1
Support_Multimodal	0.80	0.77	0.78	0.79	0.83	**0.81**
Support_Text	0.70	0.68	0.69	0.70	0.69	**0.70**
Insufficient_Multimodal	0.66	0.72	0.69	0.71	0.72	**0.73**
Insufficient_Text	0.76	0.72	**0.74**	0.74	0.72	0.73
Refute	0.96	1.00	0.98	0.99	0.98	0.98
Weighted Avg.	0.78	0.78	0.78	0.79	0.79	**0.79**

1% and 2% respectively. Final results across 7 different experiment setup shows that combining SBERT-QA at top K sentence-level evidence passage retrieval achieves optimal performance compared to the base model without ER and the use of all-round SBERT model. The best model "SBERT-QA_sentence_ER_top5" obtains 0.79 weighted average F1 at the 20th epoch.

Table 4
Factify Official Leaderboard

Rank	Team	Support_Text	Support_Multi.	Insufficient_Text	Insufficient_Multi.	Refute	Final
1	Triple-Check	**0.828**	**0.914**	0.852	**0.892**	**1.0**	**0.818**
2	INO	0.812	0.9	**0.888**	0.852	0.999	0.808
3	Logically	0.804	0.905	0.844	0.856	0.985	0.79
4	Zhang	0.766	0.879	0.816	0.879	0.999	0.774
5	gzw	0.785	0.863	0.814	0.833	1.0	0.761
6	coco	0.773	0.865	0.815	0.83	1.0	0.757
7	Noir	0.771	0.873	0.785	0.816	0.997	0.745
8	Yet	0.707	0.826	0.786	0.719	1.0	0.691
9	TeamX	0.582	0.709	0.537	0.556	0.698	0.456
-	BASELINE	0.5	0.827	0.802	0.759	0.988	0.65

6.1. Competition Result

The final test set results and competition leaderboard are presented in Table 4. The results show that top 3 participating systems achieves similar performance and our system is ranked at 3rd place with a small margin (by 0.028) to the top performing system. Please refer to [43] for the competition details.

7. Conclusion

In this research, we present our multimodal fact checking system that is submitted to the De-Factify 2023 competition. The system consists of various components, including a multimodal fact checking dataset, a QA-enhanced evidence passage retrieval component, and a Transformer-based cross-modal sequence-to-sequence veracity prediction model. Our findings from the De-Factify 2023 competition show that recent advances in pre-trained cross-modal models, such as CLIP, have strong zero-shot or few-shot capabilities and can be effectively transferred to a variety of downstream tasks, including multimodal fact checking. However, there is still a need for more effective techniques for multimodal modeling and explainability, particularly in regards to learning finer-grained cross-modal representations by jointly modeling intra- and inter-modality relationships and aligning vision regions with sentence words or entities. Additionally, more focus should be placed on real-world challenges that involve handling large amounts of textual and multimodal information from multiple sources and domains for claim verification. There is also a need for techniques that can effectively handle more complex and nuanced real-world scenarios, such as those involving sarcasm, irony, and misleading context. The difficulties in creating large and high-quality multimodal fact checking datasets that accurately reflect real-world scenarios (e.g., insufficient/leaked evidence), as identified in previous work [44, 3], remain a significant challenge.

References

[1] D. Wadden, K. Lo, L. Wang, A. Cohan, I. Beltagy, H. Hajishirzi, Multivers: Improving scientific claim verification with weak supervision and full-document context, in: Findings of the Association for Computational Linguistics: NAACL 2022, 2022, pp. 61–76.

[2] D. Stammbach, Evidence selection as a token-level prediction task, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), EMNLP, Association for Computational Linguistics (ACL), 2021.

[3] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, A. Bandhakavi, Logically at factify 2022: Multimodal fact verification, arXiv (2021). URL: https://arxiv.org/abs/2112.09253. doi:10.48550/ARXIV.2112.09253.

[4] 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.

[5] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).

[6] X. Ma, C. Zhou, X. Kong, J. He, L. Gui, G. Neubig, J. May, L. Zettlemoyer, Mega: Moving average equipped gated attention, arXiv preprint arXiv:2209.10655 (2022).

[7] J. T. Smith, A. Warrington, S. W. Linderman, Simplified state space layers for sequence modeling, arXiv preprint arXiv:2208.04933 (2022).

[8] J. Zhao, J. Bao, Y. Wang, Y. Zhou, Y. Wu, X. He, B. Zhou, Ror: Read-over-read for long document machine reading comprehension, arXiv preprint arXiv:2109.04780 (2021).

[9] R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al., Webgpt: Browser-assisted question-answering with human feedback, arXiv preprint arXiv:2112.09332 (2021).

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

[11] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: https://aclanthology.org/N18-1074. doi:10.18653/v1/N18-1074.

[12] R. Aly, Z. Guo, M. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, Feverous: Fact extraction and verification over unstructured and structured information, arXiv preprint arXiv:2106.05707 (2021).

[13] L. Floridi, M. Chiriatti, Gpt-3: Its nature, scope, limits, and consequences, Minds and Machines 30 (2020) 681–694.

[14] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, On faithfulness and factuality in abstractive summarization, arXiv preprint arXiv:2005.00661 (2020).

[15] A. Lazaridou, E. Gribovskaya, W. Stokowiec, N. Grigorev, Internet-augmented language models through few-shot prompting for open-domain question answering, arXiv preprint arXiv:2203.05115 (2022).

[16] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le,

Finetuned language models are zero-shot learners, arXiv preprint arXiv:2109.01652 (2021).

[17] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[18] 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).

[19] 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).

[20] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models, arXiv preprint arXiv:2205.12487 (2022).

[21] A. Barrón-Cedeño, T. Elsayed, P. Nakov, G. Da San Martino, M. Hasanain, R. Suwaileh, F. Haouari, N. Babulkov, B. Hamdan, A. Nikolov, et al., 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, Springer, 2020, pp. 215–236.

[22] N. Lee, C.-S. Wu, P. Fung, Improving large-scale fact-checking using decomposable attention models and lexical tagging, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1133–1138.

[23] T. Schuster, A. Fisch, R. Barzilay, Get your vitamin c! robust fact verification with contrastive evidence, arXiv preprint arXiv:2103.08541 (2021).

[24] C. Sun, A. Myers, C. Vondrick, K. Murphy, C. Schmid, Videobert: A joint model for video and language representation learning, arXiv (2019). URL: https://arxiv.org/abs/1904.01766. doi:10.48550/ARXIV.1904.01766.

[25] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, K.-W. Chang, Visualbert: A simple and performant baseline for vision and language, arXiv (2019). URL: https://arxiv.org/abs/1908.03557. doi:10.48550/ARXIV.1908.03557.

[26] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, arXiv (2019). URL: https://arxiv.org/abs/1909.11740. doi:10.48550/ARXIV.1909.11740.

[27] 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.

[28] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, Y. Qiao, Clip-adapter: Better vision-language models with feature adapters, arXiv preprint arXiv:2110.04544 (2021).

[29] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, B. Cui, Calip: Zero-shot enhancement of clip with parameter-free attention, arXiv preprint arXiv:2209.14169 (2022).

[30] K. Jiang, X. He, R. Xu, X. E. Wang, Comclip: Training-free compositional image and text matching, arXiv preprint arXiv:2211.13854 (2022).

[31] W.-Y. Wang, W.-C. Peng, Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification, arXiv (2022). URL: https://arxiv.org/abs/2201.11664. doi:10.48550/ARXIV.2201.11664.

[32] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, S. Marchand-Maillet, Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders, ACM

Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (2021) 1–23.

[33] 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.

[34] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and verification (fever) shared task, arXiv preprint arXiv:1811.10971 (2018).

[35] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fever2. 0 shared task, in: Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER), 2019, pp. 1–6.

[36] R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, A. Mittal, The fact extraction and verification over unstructured and structured information (feverous) shared task, in: Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER), 2021, pp. 1–13.

[37] D. Wadden, K. Lo, Overview and insights from the sciver shared task on scientific claim verification, arXiv preprint arXiv:2107.08188 (2021).

[38] K. Jiang, R. Pradeep, J. Lin, Exploring listwise evidence reasoning with t5 for fact verification, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2021, pp. 402–410.

[39] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, Mpnet: Masked and permuted pre-training for language understanding, Advances in Neural Information Processing Systems 33 (2020) 16857–16867.

[40] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Advances in neural information processing systems 26 (2013).

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv (2017). URL: https://arxiv.org/abs/1706.03762. doi:10.48550/ARXIV.1706.03762.

[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization (2017). URL: https://arxiv.org/abs/1711.05101. doi:10.48550/ARXIV.1711.05101.

[43] 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.

[44] M. Glockner, Y. Hou, I. Gurevych, Missing counter-evidence renders nlp fact-checking unrealistic for misinformation, arXiv preprint arXiv:2210.13865 (2022).