algorithms

MDPI

Article

Leverage Boosting and Transformer on Text-Image Matching for Cheap Fakes Detection †

Tuan-Vinh La [1,2,*], Minh-Son Dao [3,*], Duy-Dong Le [4], Kim-Phung Thai [4], Quoc-Hung Nguyen [4] and Thuy-Kieu Phan-Thi [4]

[1] University of Information Technology, Ho Chi Minh City 700000, Vietnam
[2] Vietnam National University, Ho Chi Minh City 700000, Vietnam
[3] National Institute of Information and Communications Technology, Tokyo 184-8795, Japan
[4] University of Economics, Ho Chi Minh City 700000, Vietnam
* Correspondence: vinhlt.16@grad.uit.edu.vn (T.-V.L.); dao@nict.go.jp (M.-S.D.)

† This paper is an extended version of our paper published in Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Newark, NJ, USA, 27–30 June 2022 (https://dl.acm.org/doi/abs/10.1145/3512731.3534210).

check for updates

Citation: La, T.-V.; Dao, M.-S.; Le, D.-D.; Thai, K.-P.; Nguyen, Q.-H.; Phan-Thi, T.-K. Leverage Boosting and Transformer on Text-Image Matching for Cheap Fakes Detection. *Algorithms* **2022**, *15*, 423. https://doi.org/10.3390/a15110423

Academic Editors: Xiang Zhang and Xiaoxiao Li

Received: 17 September 2022
Accepted: 11 October 2022
Published: 10 November 2022

Abstract: The explosive growth of the social media community has increased many kinds of misinformation and is attracting tremendous attention from the research community. One of the most prevalent ways of misleading news is cheapfakes. Cheapfakes utilize non-AI techniques such as unaltered images with false context news to create false news, which makes it easy and "cheap" to create and leads to an abundant amount in the social media community. Moreover, the development of deep learning also opens and invents many domains relevant to news such as fake news detection, rumour detection, fact-checking, and verification of claimed images. Nevertheless, despite the impact on and harmfulness of cheapfakes for the social community and the real world, there is little research on detecting cheapfakes in the computer science domain. It is challenging to detect misused/false/out-of-context pairs of images and captions, even with human effort, because of the complex correlation between the attached image and the veracity of the caption content. Existing research focuses mostly on training and evaluating on given dataset, which makes the proposal limited in terms of categories, semantics and situations based on the characteristics of the dataset. In this paper, to address these issues, we aimed to leverage textual semantics understanding from the large corpus and integrated with different combinations of text-image matching and image captioning methods via ANN/Transformer boosting schema to classify a triple of (image, caption$_1$, caption$_2$) into OOC (out-of-context) and NOOC (no out-of-context) labels. We customized these combinations according to various exceptional cases that we observed during data analysis. We evaluate our approach using the dataset and evaluation metrics provided by the COSMOS baseline. Compared to other methods, including the baseline, our method achieves the highest Accuracy, Recall, and F1 scores.

Keywords: deep learning; computer vision; natural language processing; image-text matching; cheapfakes; misinformation; transformer encoder

1. Introduction

In recent years, the amount of information and news has dramatically increased due to the convenience and development of social media. However, besides the benefit of its growth, it also significantly increases the quantity and impact of misinformation on individuals and society, which is one of the most dangerous things that threaten democracy, journalism, and freedom of expression. Fake news disturbs the community on a multimedia platform and causes fatal consequences in many aspects of reality and for the ordinary lives of many people. For example, fake news affected the 2016 and 2020 U.S elections.

Besides the spread of the amount of false information, the way of spreading misleading information to the community has also changed and evolved in many types and formations, making it more effective at and convenient for deceiving humans. For example, the enlargement and popularity of microblogging platforms such as Twitter, Facebook and Instagram has also increased the speed of spreading rumours and fake news since social media platforms are becoming more and more usual and necessary things in ordinary life for many people. Furthermore, controlling the content and veracity of posts on microblogging platforms is difficult since there is a large number of users on the standard platforms such as Facebook, Twitter and Instagram.

The blossoming of deep learning has opened new domains and technology, one of which is deepfake [1,2]. Deepfake has received attention from the computer vision community and is a powerful technique that can manipulate images/videos with high quality and that are hard to discriminate from unaltered ones. However, despite the usefulness and effectiveness of deepfake in swaying people's beliefs, one of the most prevalent and frequent ways of spreading disinformation is out-of-context photos, which use unaltered images in news or posts with false context.

Cheapfakes are a type of fake news that utilizes both images and new context. The danger of cheapfakes is that they are easy and cheap to make. While deepfakes use deep learning, which takes high technology and complexity to create, cheapfakes make use of simple and non-AI techniques such as photoshop, manipulating video speed, or unaltered images/videos from different events with false context, which makes it simple to create and more common.

Based on the MIT technology review (https://www.technologyreview.com/2020/12/22/1015442/cheapfakes-more-political-damage-2020-election-than-deepfakes/, accessed on 7 October 2022), in the 2020 U.S presidential election, deepfakes did not disrupt the US election, but cheapfakes did. Fazio [3] also warned of the dangers and explained why out-of-context photos are compelling. First, photos are usually attached to news, and people are already used to them. Secondly, photos make people faster at retrieving an image-related event, making it feel more truthful. Lastly, by using photos, posts on social media platforms will receive more attention and help spread false information.

To meet the emerging requirements of having a good tool for cheapfakes detection and overcome the limitations of existing works, we propose several approaches that utilize multimodal representation learning techniques to overcome limitations. By combining several techniques, including text entailment, image text matching, and boosting algorithms, our methods have improved performance and assessed the performance of several methods in cheapfakes detection.

2. Related Work

This section briefly surveys fake news detection methods, including cheapfakes detection and other subdomain methods.

2.1. Fake News Detection

Fake news has existed for a long time, even before the internet appeared. Recently, fake news has been one of the most prevalent ways to spread disinformation to human society. There are many research and public datasets on this issue. Usually, the research topic and public dataset focus on the textual type of fake news. LIAR [4] and FEVER [5] are two famous public datasets where data are collected from the news website. Each consists of one statement and a given claim, with multiple grades to determine the relation and veracity. Classification news-based linguistic semantic features [6,7] and data mining [8,9] are two traditional methods for determining the veracity of the news based on the semantics of the given text. This approach relies on training and the given data, and cannot utilize external knowledge to verify the news. Based on the development in the data and methods of the knowledge graph, Refs. [10–12] make use of the knowledge graph as external knowledge. This approach is ideal in theory, but in reality the knowledge graph suffers from a lack of relation between entities and still has a long way to develop. Although the task usually focuses on textual fake news, there are many implications for the impact on detecting disinformation in both images and text.

2.2. Rumour Detection

Alongside fake news detection, rumour detection also has a long history. Rumours refer to information not confirmed by official sources that spreads on social media platforms. Unlike fake news, which consists primarily of textual information, rumours include many types of information such as reactions, comments, attached images, user profiles, and platforms. In rumor spreading, followers play an essential role when directly or directly contribute 86 exponential increments of rumors by forwarding news with or without their comments whose content could distort the original one. Hence, understanding the following (i.e., a series of comments tailored from original news), especially in social networks, can help filter out fake news. Because data collected from social networking services can contain more attributes than data collected from news websites, such as user profiles, attached relevant posts, reactions, and comments, the data are rich and have complex attributes. The following research also has various approaches compared to fake news detection. Tree structure, sequence network [13,14] and graph neural network [15,16] are common approaches for combining and extracting correlation features on sequence and time-series data from microblogging.

2.3. Fact Checking

Fact-checking is the task of classifying the veracity of a given claim. It is a time-consuming task to verify a given claim. People need to search and check the source website's reputation and impact. Some given claims even need several professionals and several days or hours. Many techniques have been researched and developed to reduce manual fact-checking to settle this issue. There are two popular dataset types for fact-checking: the first is to verify a given pair of claims and evidence [17]. Prior research has utilized text entailment [18] to compare semantic relations between claims and evidence. Liangming et al. [19] also utilized question-answering by generating questions from the given claim. The second utilizes data on a large scale, and processes based on the technique of the knowledge graph [20].

2.4. Verify Claim about Images

Besides fake news detection, rumour detection, and fact-checking, verifying claims about an image has also received attention in recent years. While the above task mainly verifies textual claims or posts, verifying the claim about the image focuses on the post/claim/caption with the attached image. This is a challenging task since verifying the veracity of the claim itself is hard, but verifying if the attached image is related or satisfactory for concluding the truth or not is even more challenging. Refs. [21–23] extract textual captions and attached images through corresponding pre-trained models then concatenate and infer through a linear layer for classifying. La et al. [24] utilized an image–text matching method to measure correlations between captions and images. Dimitrina et al. [25] also took advantage of Google image search to enrich information (website, categories of news, and images) and then made use of TF.IDF to predict veracity.

2.5. Multi/Cross-Modal Representation Learning

In the field of multimodal reasoning and matching, many techniques have been developed to resolve various challenging tasks such as Visual Question Answering (VQA) [26], Image Captioning [27], Text-to-Image [28], and Image–Text Matching [29]. Still, there is much research on the cross-modal between images and text. To verify claims about image tasks, many methods use the simple technique of extracting features of images through Convolution Neural Network and concatenating them with textual features to classify the truthfulness of news. This technique is simple yet depends on the training dataset, which cannot be generalized in reality and for other aspects and types of news.

3. Dataset

This section will briefly introduce the Out-of-Context Detection Dataset in COSMOS [30], which we used to assess and evaluate our proposal's performance. The dataset was collected from news websites (New York Times, CNN, Reuters, ABC, PBS, NBCLA, AP News, Sky News, Telegraph, Time, DenverPost, Washington Post, CBC News, Guardian, Herald Sun, Independent, CS Gazette, BBC) and fact-checking websites. The dataset consisted of the English language in 19 categories and did not consist of digitally-altered or fake images. The statistic is shown in Figure 1 and Table 1. We recommend readers read [31] for more details.

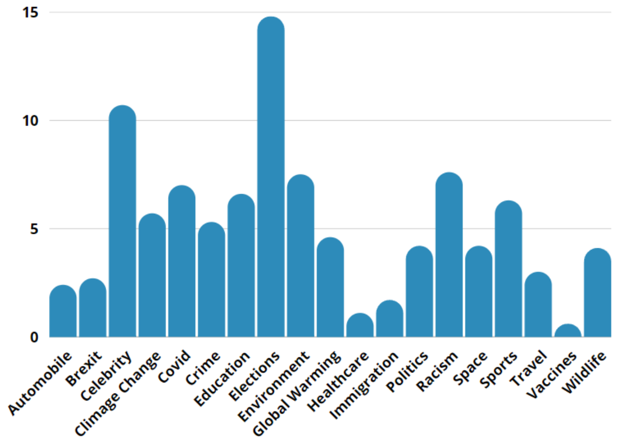
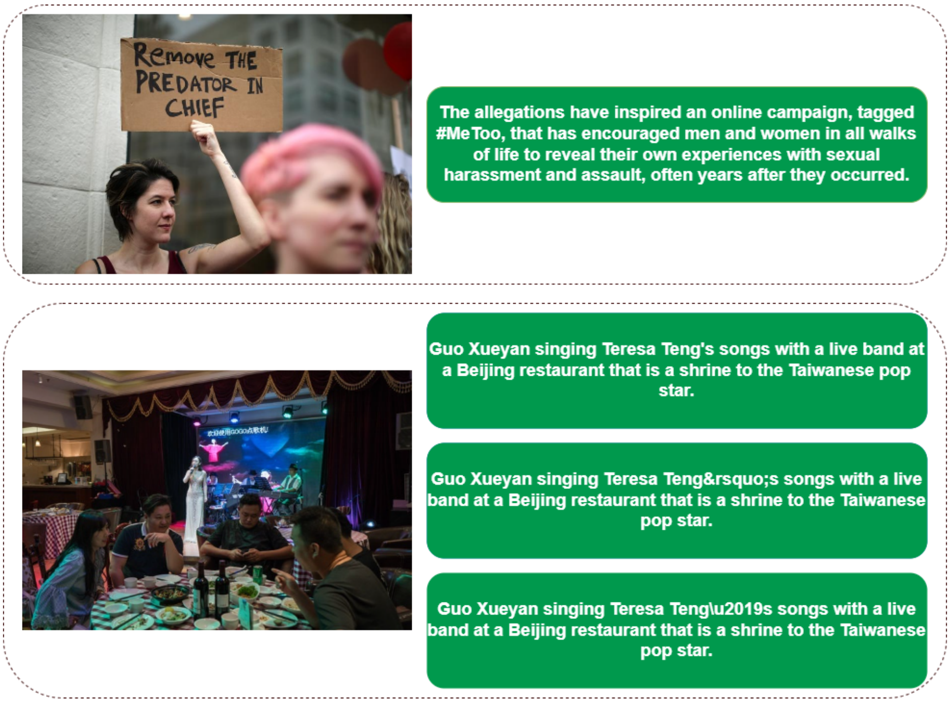


Figure 1. Distribution in categories and content of COSMOS dataset.

Table 1. COSMOS Dataset statistic.

Dataset	Images	Captions	Context Annotation
Training	161,752	360,749	✗
Validation	41,006	90,036	✗
Public Test	1000	2000	✓

Train/Validate Set: In the training set, captioned images were collected from the news website. Each captioned image consisted of one image, one or multiple attached captions, source URL, entity list in a caption, modified caption in which each entity is replaced by corresponding ontology, and location of 10 bounding boxes extracted by a pre-trained Mask-RCNN on MS COCO. Training data did not contain an out-of-context captioned image. Every captioned image was not-out-of-context and did not have a context label. Training data consisted of around 200,000 images with 450,000 matching textual captions. Furthermore, 20% of that was split for the validation set. The example of the captioned image of the training set is illustrated in Figure 2.



The allegations have inspired an online campaign, tagged #MeToo, that has encouraged men and women in all walks of life to reveal their own experiences with sexual harassment and assault, often years after they occurred.

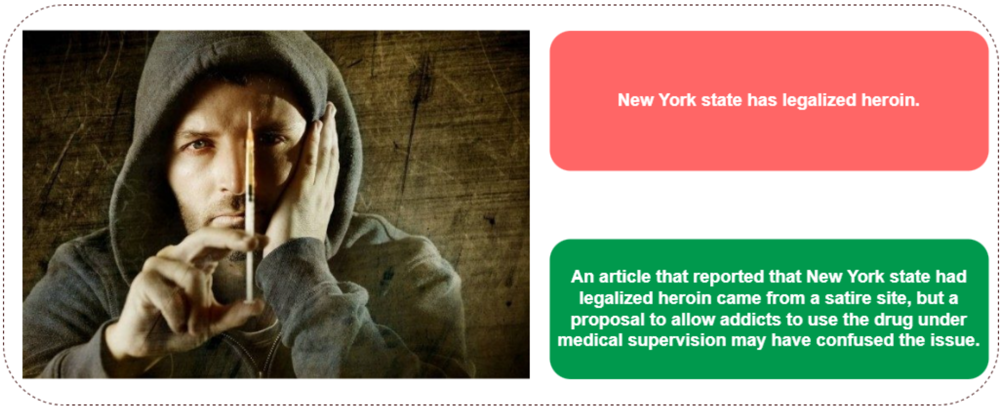
Guo Xueyan singing Teresa Teng's songs with a live band at a Beijing restaurant that is a shrine to the Taiwanese pop star.

Guo Xueyan singing Teresa Teng&rsquo;s songs with a live band at a Beijing restaurant that is a shrine to the Taiwanese pop star.

Guo Xueyan singing Teresa Teng\u2019s songs with a live band at a Beijing restaurant that is a shrine to the Taiwanese pop star.

Figure 2. Example of the captioned image in the training set. Training data do not contain an out-of-context captioned image. Every captioned image is not-out-of-context and does not have a context label.

Test Set: In the test set, captioned images were collected from both news websites and fact-checking websites. Like the train set, each captioned image of the test set consisted of an image, captions, source URL, entity list, modified caption, and bounding box. However, each captioned image contained two corresponding captions in the test set. These captions always contained one caption not-out-of-context; the remaining caption could be out-of-context or not-out-of-context. Each captioned image also had context annotation to point out if that captioned image consisted of out-of-context captions or not. In summary, the test set contained 1000 captioned images, which included 1000 images and 2000 textual captions. The example of the captioned image of the test set is illustrated in Figure 3.



New York state has legalized heroin.

An article that reported that New York state had legalized heroin came from a satire site, but a proposal to allow addicts to use the drug under medical supervision may have confused the issue.

Figure 3. Example of the captioned image in the testing set. The captioned image contains one image and two corresponding captions. These captions always have one caption not-out-of-context; the remaining caption can be out-of-context or not-out-of-context.

4. Proposed Method

In this section, we will introduce COSMOS baseline [30], our motivation, and explain and describe our methods.

4.1. COSMOS Baseline

In prior research on image and news veracity classification, the method usually aims to utilize multi-modal by extracting features of text/captions and attached images through a pre-trained convolution neural network, LSTM [32] or BERT [33], layer and combine these features by concatenating or sum function with the appropriate objective function. This approach can take advantage of multiple datasets such as imagenet, MSCOCO, STS, and MNLI... for the basis of understanding and representing semantic information of data and fine-tuning other news datasets to improve performance.

Besides the advantage of prior research, it is also limited in terms of the dataset's attributes. Most of the prior work uses fine-tuning on the new dataset, which makes it limited in many respects, such as in categories and characteristics of news, and cannot cover all subjects or situations not included in the dataset.

In COSMOS, the author aims to match the caption with the most correlated object in the image by utilizing self-supervised learning. To do this, the author first uses Mask-RCNN [34] on MSCOCO [35] and selects the top 10 ROIs (Region of Interest) with the highest detection score and additional features of the entire image. For text pre-processing and processing, the author first makes use of NER (Named Entity Recognition) to generalize captions and then infers through USE (Universal Sentence Encoder) [36] to extract caption embedding. Next, the author infers the bounding box and caption embedding through a linear layer for mapping to the same dimension. The paper also uses max margin ranking loss [30] as objective/loss function using the equation:

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\max(0, (S^{r}_{IC} - S^{m}_{IC}) + \alpha), \tag{1}$$

where S^{r}_{IC}, S^{m}_{IC} is the measure of similarity between a random caption–image pair and a matching caption–image pair, and α is the margin parameter. This measure is calculated by the maximum dot function between 11 ROIs and matching/random caption. The similarity measure function is illustrated as Equation :

$$S_{IC} = \max_{i}^{N} (b_i^T c), \tag{2}$$

where b_i is the features of the proposal bounding box and c is the features of the caption.

At testing time, for each captioned image (caption$_1$,caption$_2$,image), the COSMOS method uses the simple if else rule to determine out-of-context captioned images:

$$\begin{cases} OOC, & \text{If IoU}(B_{IC_1}, B_{IC_2}) > t_i \;\&\; S_{sim}(C_1, C_2) < t_c \\ NOOC, \text{otherwise}, \end{cases} \tag{3}$$

where IoU(B_{IC_1}, B_{IC_2}) is the intersection-over-union of two bounding boxes having the largest value of similarity measure with the corresponding two captions; $S_{sim}(C_1, C_2)$ is the similarity measure defined in cosine space, and t_i, t_c is the fixed threshold of IoU(B_{IC_1}, B_{IC_2}) and $S_{sim}(C_1, C_2)$.

By matching and comparing two captions with the corresponding object, the author can assess if two captions mention a related subject/object or not (determined by IoU(B_{IC_1}, B_{IC_2})). If two captions mention a related subject/object and have uncorrelated semantic similarity (determined by $S_{sim}(C_1, C_2)$), then the given captioned image is out-of-context. The other situation is not-out-of-context.

4.2. Motivation

By training the model matching caption with the correlated object in the image and utilizing a pre-trained large-scale textual dataset, the method can utilize the semantic features and understanding of another large-scale dataset, which make it less prone to overfitting on other tasks or datasets of news or fact verification.

Besides the advantages of the COSMOS baseline, the weakness of this method is that by utilizing features of the entire image of Mask-RCNN on MSCOCO, it cannot optimize the express context of the entire image because the Mask-RCNN's task is object detection, not descripting. Moreover, the caption usually mentions multiple objects and highly correlates with the context image.

Based on the insufficiency of the COSMOS method when comparing the image with the caption, in this paper, we propose and evaluate a method that utilizes a more optimized method to express content features of the image and better extracts the semantic relation between two captions. Furthermore, instead of defining a rule for determining out-of-context captioned images, we combined results from multiple methods by making use of boosting techniques to improve performance.

4.3. Methodology

This paper proposes two approaches to measuring the correlation between image and caption: image captioning and image–caption matching.

Image Captioning: For the image captioning approach, we aim to utilize [37] to generate the content description of an image. We can use a pre-trained large-scale dataset on the STS [38] task (Semantic Textual Similarity) to measure the correlation between caption and image by converting the image's content to textual form.

Image-Caption Matching: For the image-caption matching approach, we utilized a trained model of image–text matching on the MSCOCO dataset [35] to measure the correlation between caption and image. In this paper, we used the Visual Semantic Reasoning [39] method to measure the similarity between image and caption. See Figure 4 for illustration.

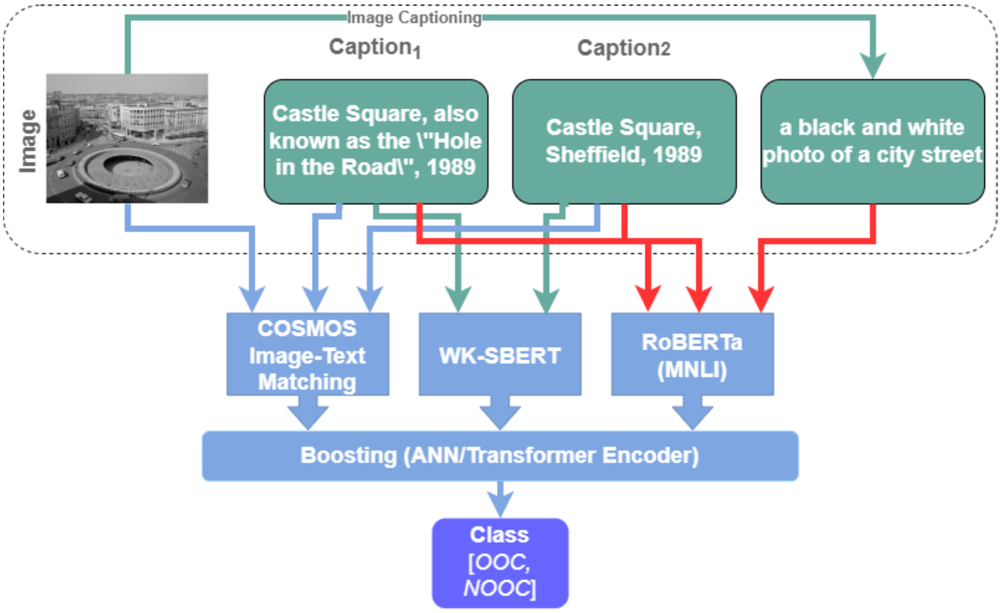


Image Captioning
Caption1
Caption2
Image
Castle Square, also known as the \"Hole in the Road\", 1989
Castle Square, Sheffield, 1989
a black and white photo of a city street
COSMOS Image-Text Matching
WK-SBERT
RoBERTa (MNLI)
Boosting (ANN/Transformer Encoder)
Class [OOC, NOOC]

Figure 4. Illustration of boosting with image captioning method. First, the image will be inferred self-critically [37] to obtain a description of the image in textual form. Next, RoBERTa(MNLI) is utilized to extract the correlation between caption$_1$, caption$_2$, and image (NLI(caption$_1$, caption$_2$), NLI(caption$_1$, caption$_{image}$), NLI(caption$_2$, caption$_{image}$)).To overcome the difference between training data and testing data issues and improve performance, we take advantage of the boosting algorithm on the part of the testing data to combine results from our proposal and the COSMOS baseline.

The VSRN (Visual Semantic Reasoning) [39] method utilizes margin ranking loss as the objective function. The margin ranking loss objective is the correlation measurement of the matching caption–image, which is higher than the non-matching caption-image and not trying to make matching caption–image have a matching score higher than the threshold. As shown in Figure 5, the matching caption image's matching score has a different range of values. It can have a lower value compared with different captions and images that do not match each other. However, compared to the same image with another caption that is not matching, the correlation measurement of the matching caption image is higher than that of the non-matching caption image. Based on this attribute of the VSRN method and margin ranking loss, we normalized the matching score using Equation (4) to overcome this issue. See Figure 6 for illustration.

$$\hat{S}(I,C) = S(I,C) - \frac{1}{2N}\sum_{r}^{N}[S(I,C_r) + S(I_r,C)], \tag{4}$$

where $\hat{S}$ defines the normalize matching score, and r defines the random index that satisfies $C_r \neq C$ and $I_r \neq I$. By subtracting the mean of the matching score from the N sample, the result can express the correlation degree of the given matching image caption compared with other non-matching image captions.

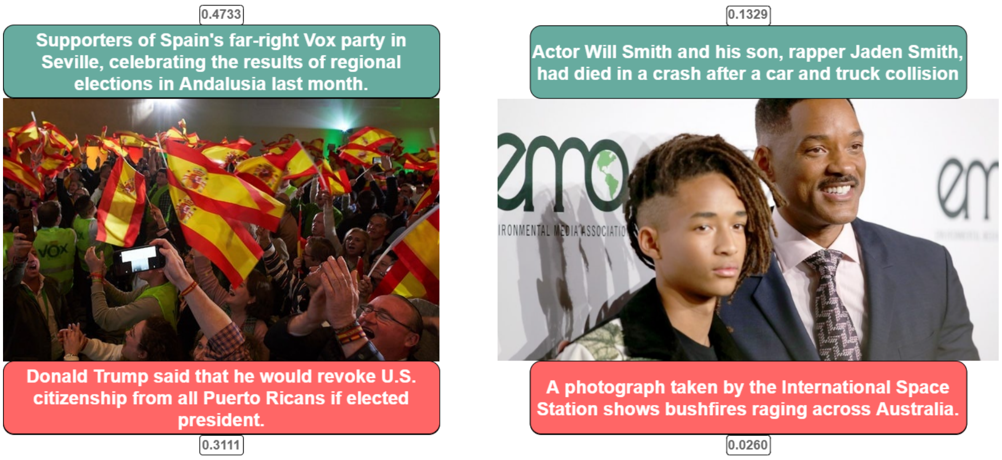

0.4733
Supporters of Spain's far-right Vox party in Seville, celebrating the results of regional elections in Andalusia last month.
Donald Trump said that he would revoke U.S. citizenship from all Puerto Ricans if elected president.
0.3111
0.1329
Actor Will Smith and his son, rapper Jaden Smith, had died in a crash after a car and truck collision
A photograph taken by the International Space Station shows bushfires raging across Australia.
0.0260

Figure 5. Example of the matching score between image and caption. Green expresses matching caption and red expresses non-matching caption. Based on the attribute of margin ranking loss, compared to one image, matching captions have a higher score than the non-matching caption. Not every matching caption always has a higher matching score than a non-matching caption.

Hence, to estimate the correlation between two captions better, instead of using only cosine similarity measures from other methods trained on the STS task [38], we also used other methods on the NLI task (Natural Language Inference) [40] to express the semantic relation between two captions. We chose SBERT-WK [41] and RoBERTa [42] to extract semantic relations between two captions.

One of the difficulties of the COSMOS dataset is that training/validation data have a different construct from testing data. In training data, each captioned image consists of only a not-out-of-context pair, and captions are always trustworthy news and match the image's context. While in testing data, data consist of out-of-context and not-out-of-context captioned images. The caption can be fake news, descriptions about the image, or match/mismatch with the image and other captions. Based on our experience, fine-tuning training data and evaluating directly on testing data gave poor results. We used boosting algorithms—which can utilize results from textual entailment (NLI, STS) and image–caption matching (image–text matching, image captioning) to increase the method's accuracy—on the part of the testing dataset to combine semantics understanding from multiple methods to improve performance and overcome the shift domain issue. We leveraged ANN and Transformer Encoder as boosting architecture. Six hundred captioned images were extracted as training data and 400 captioned images as evaluation data.

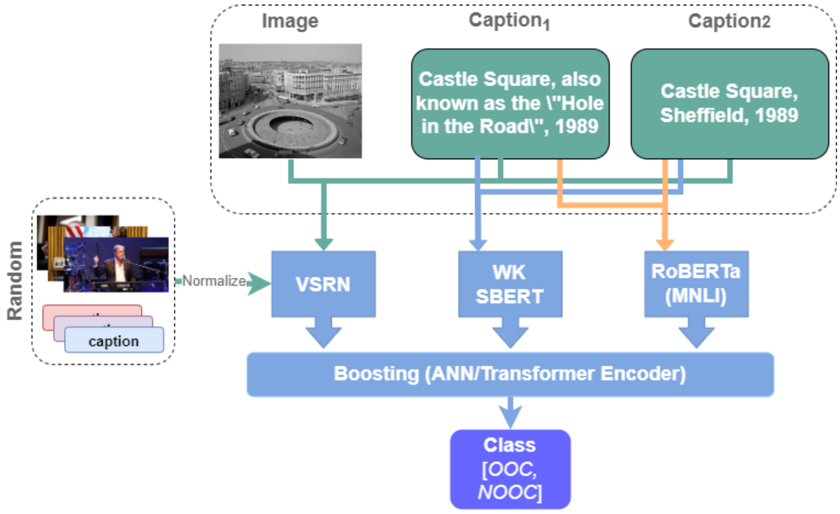


Image
Caption1
Caption2
Castle Square, also known as the \"Hole in the Road\", 1989
Castle Square, Sheffield, 1989
Random
caption
Normalize
VSRN
WK SBERT
RoBERTa (MNLI)
Boosting (ANN/Transformer Encoder)
Class [OOC, NOOC]

Figure 6. Illustration of boosting with image–caption matching method. First, image, caption$_1$, and caption$_2$ will be inferred through VSRN [39] and normalized by Equation (4) to obtain matching scores ($\hat{\mathrm{S}}(I, C_1), \hat{\mathrm{S}}(I, C_2)$). In addition to enriching semantic correlation information between caption$_1$ and caption$_2$, we make use of RoBERTa(MNLI) to extract the relation between two captions. Similar to the image captioning method, we take advantage of the boosting algorithm on the part of testing data to combine results from our proposal and the COSMOS baseline.

We also used a boosting algorithms on a combination of mixed results to compare the effects of each component. In summary, we evaluated the performance of boosting algorithms on a set of components:

- Boosting combination of $\mathrm{IoU}(B_{IC_1}, B_{IC_2})$ and $S_{sim}(C_1, C_2)$ using ANN;
- Boosting combination of $\mathrm{IoU}(B_{IC_1}, B_{IC_2})$, $S_{sim}(C_1, C_2)$, $NLI(C_1, C_2, C_{image})$ using ANN [43];
- Boosting combination of $\mathrm{IoU}(B_{IC_1}, B_{IC_2})$ and $S_{sim}(C_1, C_2)$ using Transformers Encoder;
- Boosting with $\mathrm{IoU}(B_{IC_1}, B_{IC_2})$, $S_{sim}(C_1, C_2)$, $NLI(C_1, C_2, C_{image})$ using Transformers Encoder;
- Boosting combination of $\mathrm{IoU}(B_{IC_1}, B_{IC_2})$, $S_{sim}(C_1, C_2)$, $\hat{\mathrm{S}}(I, C_1)$, $\hat{\mathrm{S}}(I, C_2)$, $NLI(C_1, C_2)$ using ANN;
- Boosting combination of $\mathrm{IoU}(B_{IC_1}, B_{IC_2})$, $S_{sim}(C_1, C_2)$, $\hat{\mathrm{S}}(I, C_1)$, $\hat{\mathrm{S}}(I, C_2)$, $NLI(C_1, C_2)$ using Transformers Encoder,

where $NLI(C_1, C_2)$ and $NLI(C_1, C_2, C_{image})$ is the result of RoBERTa [42] on the NLI task given three pairs of sentences (C_1, C_2), (C_1, C_{image}), (C_2, C_{image}). The result contains three probabilities of three class that express the semantic relationship between two captions/sentences: entailment, neutral, and contradiction. We illustrated an example of boosting with image captioning and image–text matching in Figures 4 and 6.

5. Experimental & Results

This section introduces the dataset and metric used to evaluate our proposed method. We compare our method to others on the same dataset and metric. The thoughtful discussion also raises the advantages and disadvantages of our method.

5.1. Working Environment

All our experimental methods were implemented on three GPUs NVIDIA Tesla A100 40 GB, Intel Xeon Gold 5220R CPU, and 256 GB RAM. We extracted 600 captioned images of testing data for boosting and 400 captioned images for evaluating performance.

We used the same settings to make it easy to compare each method's performance. We used an Adam optimizer with a 1×10^{-3} learning rate, 4×10^{-5} weight decay, and cross-entropy loss for an updated model. We used simple ANN and a Transformers Encoder to boost the results.

We set the default target dimension for ANN to 64, fed-forward the activation layer (PReLU), and inferred through the linear layer to classify the captioned image.

For the Transformers Encoder, we set input features to 16 dimensions, two multi-head attention, and two layers to extract features. After that, we inferred through the linear layer to classify the captioned image.

5.2. Evaluation Metrics

To evaluate the effectiveness of our proposal, we used five metrics: accuracy, precision, recall, and F1-score with the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1score = \frac{2 \times Recall \times Precision}{Recall \times Precision}, \tag{8}$$

where:

- True Positives (TP): Number of samples correctly identified as out-of-context;
- True Negatives (TN): Number of samples correctly identified as not-out-of-context;
- False Positives (FP): Number of samples incorrectly identified as out-of-context;
- False Negatives (FN): Number of samples incorrectly identified as not-out-of-context.

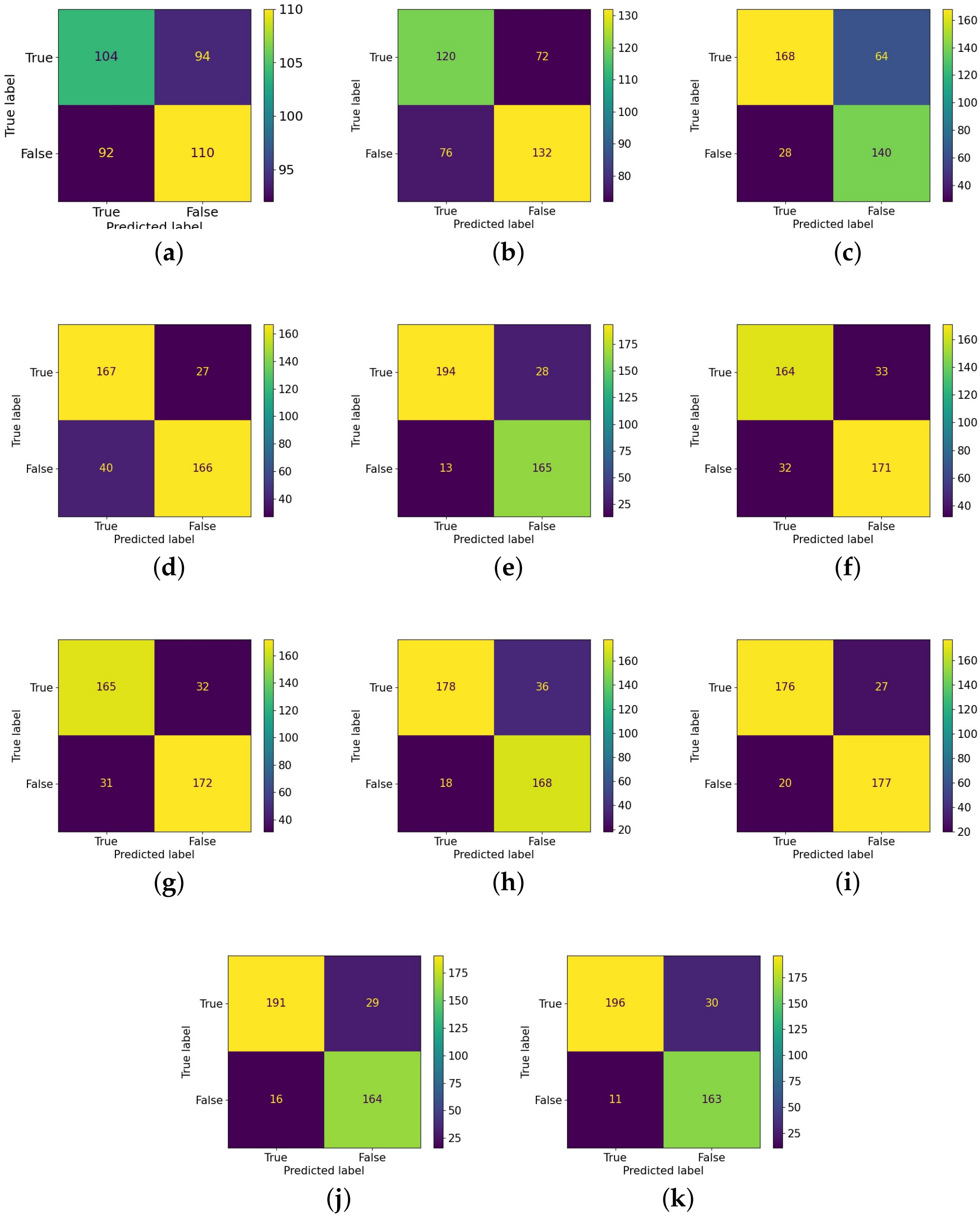
5.3. Datasets and Compared Methods

We evaluated our proposals and other methods on 400 captioned image testing datasets. Table 2 and Figure 7 summarize the result of our proposal compared with other methods.

Table 2. The Comparisons.

Method	Accuracy	Precision	Recall	F1-Score
Spotfake [21]	0.535	0.5252	0.5306	0.5279
EANN [21]	0.63	0.6025	0.6122	0.6185
SBERT-WK [41]	0.77	0.7241	0.8571	0.7850
COSMOS Baseline [30]	0.8325	0.8608	0.8067	0.8329
Tankut et al. [44]	**0.8975**	**0.8738**	0.9371	0.9044
Boosting with IoU(B_{IC_1}, B_{IC_2}) and $S_{sim}(C_1, C_2)$ with ANN	0.8375	0.8324	0.8367	0.8346
Boosting with IoU(B_{IC_1}, B_{IC_2}) and $S_{sim}(C_1, C_2)$ with Transformers Encoder	0.8425	0.8375	0.8418	0.8396
Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$, $NLI(C_1, C_2, C_{image})$ with ANN [43]	0.865	0.8317	0.9081	0.8682
Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$, $NLI(C_1, C_2, C_{image})$ with Transformers Encoder	0.8825	0.8669	0.8979	0.8822
Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$, $\hat{S}(I, C_1), \hat{S}(I, C_2)$, & $NLI(C_1, C_2)$ with ANN	0.8875	0.8681	0.9227	0.8946
Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$, $\hat{S}(I, C_1), \hat{S}(I, C_2)$ & $NLI(C_1, C_2)$ with Transformers Encoder	**0.8975**	0.8672	**0.9468**	**0.9053**

Bold factor meaning best evaluation score.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

(**g**) (**h**) (**i**)

(**j**) (**k**)

Figure 7. Confusion matrix of (**a**)Spotfake [23]; (**b**)EANN [21]; (**c**) SBERT-WK [41]; (**d**) COSMOS Baseline [30]; (**e**) COSMOS on Steroid [44]; (**f**) Boosting IoU(B_{IC_1}, B_{IC_2}) $S_{sim}(C_1, C_2)$ with ANN; (**g**) Boosting IoU(B_{IC_1}, B_{IC_2}) & $S_{sim}(C_1, C_2)$ with Transformers Encoder; (**h**) Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$ & $NLI(C_1, C_2, C_{image})$ with ANN; (**i**) Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$ & $NLI(C_1, C_2, C_{image})$ with Transformers Encoder; (**j**) Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$ $\hat{S}(I, C_1)$, $\hat{S}(I, C_2)$ & $NLI(C_1, C_2)$ with ANN; (**k**) Boosting with IoU(B_{IC_1}, B_{IC_2}), $S_{sim}(C_1, C_2)$ $\hat{S}(I, C_1)$, $\hat{S}(I, C_2)$ & $NLI(C_1, C_2)$ with Transformers Encoder.

5.4. Discussions

First, we made use of Spotfake [23] as a training baseline approach based on its simplicity—fine-tuning and concatenating visual and textual embedding to classify the veracity of the news. We leveraged Spotfake architecture on the given training and testing data of COSMOS. In particular, when training, we created out-of-context content by selecting captions and images from different sources' captioned images and not-out-of-context content from the same source captioned images. When evaluating, we classified both (caption$_1$, image) and (caption$_2$, image). If both the captions were not-out-of-context, the triplet (caption$_1$, caption$_2$, image) was not-of-context, and the other was out-of-context. The method gave poor results based on the different attributes between training and testing data, and the method could not overcome and generalize the issue.

Next, downstream from another dataset approach, we chose EANN. We used the same method from Spotfake to evaluate the performance—classify both (caption$_1$, image) and (caption$_2$, image). On the MediaEval2015 dataset [45], EANN could achieve a 71.5% accuracy point. However, when downstream of COSMOS, the method produced unqualified results, even though MediaEval2015 consists of a large corpus of textual news and various cases of misused images, similar to the COSMOS dataset. The current training and downstream approach to a given news dataset is limited in categories, domains, and types of news and may not perform well in reality.

Compared to the baseline, our methods improved the 6.5% accuracy score. Furthermore, in relation to Tankut et al.'s [44] research, our method has equal accuracy and has a higher recall and F1-score. Tankut et al. [44] took advantage of handcraft features by matching the most relevant fake news keywords (fake, hoax, fabrication, supposedly, falsification, propaganda, deflection, deception, contradiction, defamation, lie, misleading, deceive, fraud, concocted, bluffing, made up, double meaning, alternative facts, tricks, half-truths, untruth, falsehoods, inaccurate, disinformation, misconception) and alternated captions in testing datasets with fake words ("was true" and "was not true") to compare semantic features. Our methods used various semantic understandings in computer vision and natural language processing on large-scale datasets to assess the correlation between the original image and caption. The impact of each image-text matching method is also present in our paper.

In Figures 8 and 9 we show a few examples of our false negative (FP) and our false positive (FN) predictions. As we can see in the false negative cases, the content of news and the abstract relation with the corresponding image are hard to distinguish, even by humans, and much news needs an expert or time with search tools to determine. For false positive cases, our method failed to distinguish between the image description (generated by humans) and false news.

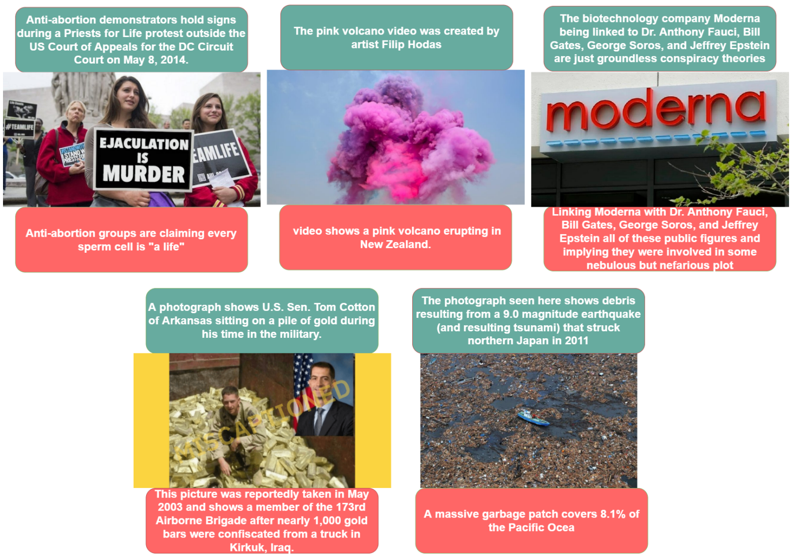

Anti-abortion demonstrators hold signs during a Priests for Life protest outside the US Court of Appeals for the DC Circuit Court on May 8, 2014.
The pink volcano video was created by artist Filip Hodas
The biotechnology company Moderna being linked to Dr. Anthony Fauci, Bill Gates, George Soros, and Jeffrey Epstein are just groundless conspiracy theories
EJACULATION IS MURDER
#TEAMLIFE
moderna
Anti-abortion groups are claiming every sperm cell is "a life"
video shows a pink volcano erupting in New Zealand.
Linking Moderna with Dr. Anthony Fauci, Bill Gates, George Soros, and Jeffrey Epstein all of these public figures and implying they were involved in some nebulous but nefarious plot
A photograph shows U.S. Sen. Tom Cotton of Arkansas sitting on a pile of gold during his time in the military.
The photograph seen here shows debris resulting from a 9.0 magnitude earthquake (and resulting tsunami) that struck northern Japan in 2011
MISCAPTIONED
This picture was reportedly taken in May 2003 and shows a member of the 173rd Airborne Brigade after nearly 1,000 gold bars were confiscated from a truck in Kirkuk, Iraq.
A massive garbage patch covers 8.1% of the Pacific Ocea

Figure 8. False negative cases. Out-of-context captioned image is classified as not-out-of-context.

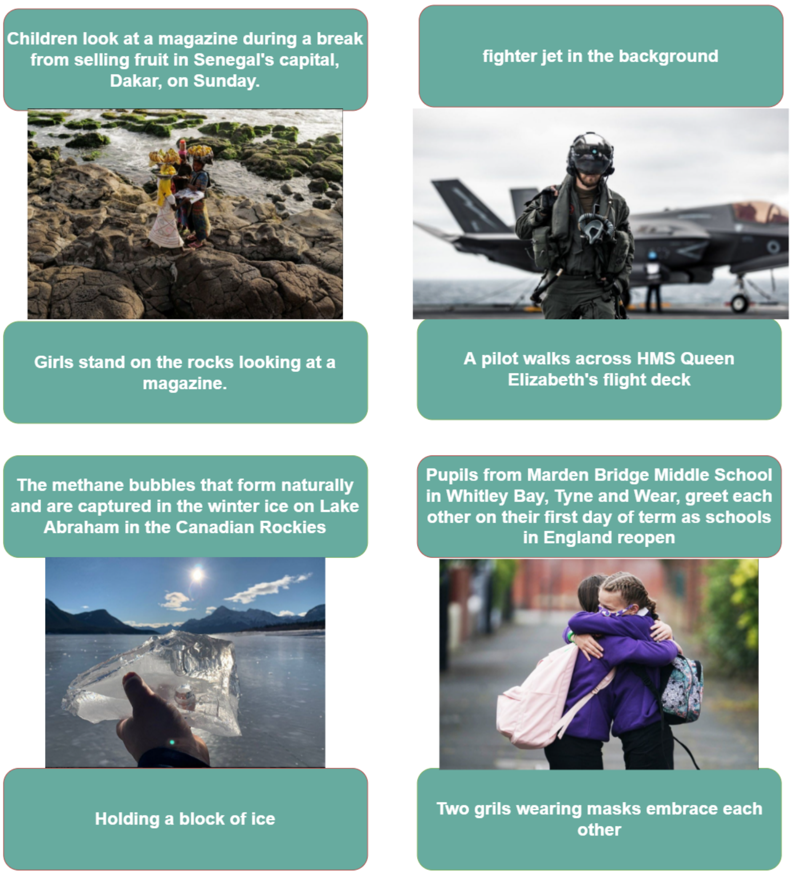

Children look at a magazine during a break from selling fruit in Senegal's capital, Dakar, on Sunday.
Girls stand on the rocks looking at a magazine.
fighter jet in the background
A pilot walks across HMS Queen Elizabeth's flight deck
The methane bubbles that form naturally and are captured in the winter ice on Lake Abraham in the Canadian Rockies
Holding a block of ice
Pupils from Marden Bridge Middle School in Whitley Bay, Tyne and Wear, greet each other on their first day of term as schools in England reopen
Two grils wearing masks embrace each other

Figure 9. False positive cases. Not-out-of-context captioned image is classified as out-of-context.

6. Conclusions

We have presented and evaluated multiple approaches to the cheapfakes detection problem and conducted experiments on the COSMOS dataset. Our work evaluates the effectiveness of different image–text matching methods, which can leverage semantic features from large-scale datasets instead of fine-tuning and concatenating features from text and images, which makes methods limited in the attribute of a given dataset. Compared to the existing method for cheapfakes detection, we have proposed a method that takes advantage of attributes from the testing dataset instead of directly alternating and defines handcraft patterns based on human effort. Moreover, we have extended experiments of the same theoretical results previously described [43]. Compared to another approach, our methods achieve competitive results, which achieve equal accuracy and higher recall and F1-score. Overall, we believe that our method makes a valuable contribution towards addressing misinformation in news and social media.

In the future, we will consider abstract images that cannot explain or understand with popular image understanding methods without specific knowledge, such as a photo of an art painter, a personal event, a snapshot from a film, or a photo of a book cover. We also consider mapping images and captions into the third coordinator, where additional knowledge can bridge the semantic/knowledge gap between them. Not but not least, extending captions using domain knowledge (e.g., hugging face) to enrich the semantic content of captions and utilize content graphs extracted from images can be another promising research direction.

Author Contributions: Project administration, M.-S.D.; conceptualization, M.-S.D.; writing—review and editing, T.-V.L.; writing—original draft preparation, T.-V.L.; methodology, T.-V.L.; formal analysis, T.-V.L.; validation, T.-V.L.; software, T.-V.L.; funding acquisition, D.-D.L.; data curation, D.-D.L., K.-P.T., Q.-H.N. and T.-K.P.-T.; resource D.-D.L., K.-P.T., Q.-H.N. and T.-K.P.-T. All authors have read and agreed to the published version of the manuscript.

Funding: This research is funded by the University of Economic Ho Chi Minh City (UEH) Vietnam grant number 2022-09-09-1144.

Data Availability Statement: The authors will make the data used in this research available on request.

Acknowledgments: We acknowledge the University of Economic Ho Chi Minh City (UEH) for funding this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Westerlund, M. The emergence of deepfake technology: A review. *Technol. Innov. Manag. Rev.* **2019**, *9*, 40–53. [CrossRef]
2. Collins, A. *Forged Authenticity: Governing Deepfake Risks*; Technical Report; EPFL International Risk Governance Center (IRGC): Lausanne, Switzerland, 2019.
3. Fazio, L. Out-of-Context Photos Are a Powerful Low-Tech Form of Misinformation. Available online: https://mat.miracosta.edu/mat210_cotnoir/instructor/pdfs-for-class/Out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation.pdf (accessed on 7 October 2022).
4. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; Mittal, A. Fever: A large-scale dataset for fact extraction and verification. *arXiv* **2018**, arXiv:1803.05355.
5. Wang, W.Y. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv* **2017**, arXiv:1705.00648.
6. Choudhary, A.; Arora, A. Linguistic feature based learning model for fake news detection and classification. *Expert Syst. Appl.* **2021**, *169*, 114171. [CrossRef]
7. Singh, V.; Dasgupta, R.; Sonagra, D.; Raman, K.; Ghosh, I. Automated fake news detection using linguistic analysis and machine learning. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS), Washington, DC, USA, 5–8 July 2017; pp. 1–3.
8. Shu, K.; Sliva, A.; Wang, S.; Tang, J.; Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explor. Newsl.* **2017**, *19*, 22–36. [CrossRef]
9. Bharadwaj, P.; Shao, Z. Fake news detection with semantic features and text mining. *Int. J. Nat. Lang. Comput. (IJNLC)* **2019**, *8*, 17–22. [CrossRef]
10. Pan, J.Z.; Pavlova, S.; Li, C.; Li, N.; Li, Y.; Liu, J. Content based fake news detection using knowledge graphs. In Proceedings of the International Semantic Web Conference, Monterey, CA, USA, 8–12 October 2018; pp. 669–683.
11. Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; Zhou, M. 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), Online, 1–6 August 2021; pp. 754–763.
12. Wang, Y.; Qian, S.; Hu, J.; Fang, Q.; Xu, C. Fake news detection via knowledge-driven multimodal graph convolutional networks. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 26–29 October 2020; pp. 540–547.
13. Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B.J.; Wong, K.F.; Cha, M. Detecting rumors from microblogs with recurrent neural networks. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, NY, USA, 9–15 July 2016.
14. Ma, J.; Gao, W.; Wong, K.F. *Rumor Detection on Twitter with Tree-Structured Recursive Neural Networks*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018.
15. Wu, Z.; Pi, D.; Chen, J.; Xie, M.; Cao, J. Rumor detection based on propagation graph neural network with attention mechanism. *Expert Syst. Appl.* **2020**, *158*, 113595. [CrossRef]
16. Bian, T.; Xiao, X.; Xu, T.; Zhao, P.; Huang, W.; Rong, Y.; Huang, J. Rumor detection on social media with bi-directional graph convolutional networks. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 549–556.
17. Mishra, S.; Suryavardan, S.; Bhaskar, A.; Chopra, P.; Reganti, A.; Patwa, P.; Das, A.; Chakraborty, T.; Sheth, A.; Ekbal, A.; 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), Vancouver, BC, Canada, 22 Februrary–1 March 2022.
18. Gao, J.; Hoffmann, H.F.; Oikonomou, S.; Kiskovski, D.; Bandhakavi, A. Logically at the factify 2022: Multimodal fact verification. *arXiv* **2021**, arXiv:2112.09253.
19. Pan, L.; Chen, W.; Xiong, W.; Kan, M.Y.; Wang, W.Y. Zero-shot fact verification by claim generation. *arXiv* **2021**, arXiv:2105.14682.
20. Ciampaglia, G.L.; Shiralkar, P.; Rocha, L.M.; Bollen, J.; Menczer, F.; Flammini, A. Computational fact checking from knowledge networks. *PLoS ONE* **2015**, *10*, e0128193.
21. Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; Gao, J. Eann: Event adversarial neural networks for multi-modal fake news detection. In Proceedings of the 24th ACM Sigkdd International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; pp. 849–857.

22. Khattar, D.; Goud, J.S.; Gupta, M.; Varma, V. Mvae: Multimodal variational autoencoder for fake news detection. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13 May–17 May 2019; pp. 2915–2921.
23. Singhal, S.; Shah, R.R.; Chakraborty, T.; Kumaraguru, P.; Satoh, S. Spotfake: A multi-modal framework for fake news detection. In Proceedings of the 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM), Singapore, 11–13 September 2019; pp. 39–47.
24. La, T.V.; Dao, M.S.; Tran, Q.T.; Tran, T.P.; Tran, A.D.; Nguyen, D.T.D. A Combination of Visual-Semantic Reasoning and Text Entailment-based Boosting Algorithm for Cheapfake Detection. In Proceedings of the ACM MM 2022, Lisbon, Portugal, 10–14 October 2022.
25. Zlatkova, D.; Nakov, P.; Koychev, I. Fact-checking meets fauxtography: Verifying claims about images. *arXiv* **2019**, arXiv:1908.11722.
26. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
27. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 121–137.
28. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual Event, 18–24 July 2021; pp. 8821–8831.
29. Lee, K.H.; Chen, X.; Hua, G.; Hu, H.; He, X. Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 201–216.
30. Aneja, S.; Bregler, C.; Nießner, M. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv* **2021**, arXiv:2101.06278.
31. Aneja, S.; Midoglu, C.; Dang-Nguyen, D.T.; Khan, S.A.; Riegler, M.; Halvorsen, P.; Bregler, C.; Adsumilli, B. ACM Multimedia Grand Challenge on Detecting Cheapfakes. *arXiv* **2022**, arXiv:2207.14534.
32. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
33. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
34. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
35. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
36. Cer, D.; Yang, Y.; Kong, S.y.; Hua, N.; Limtiaco, N.; John, R.S.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; et al. Universal sentence encoder. *arXiv* **2018**, arXiv:1803.11175.
37. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
38. Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; Specia, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv* **2017**, arXiv:1708.00055.
39. Li, K.; Zhang, Y.; Li, K.; Li, Y.; Fu, Y. Visual semantic reasoning for image-text matching. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 4654–4662.
40. Williams, A.; Nangia, N.; Bowman, S.R. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv* **2017**, arXiv:1704.05426.
41. Wang, B.; Kuo, C.C.J. Sbert-wk: A sentence embedding method by dissecting bert-based word models. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2146–2157. [CrossRef]
42. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
43. La, T.V.; Tran, Q.T.; Tran, T.P.; Tran, A.D.; Dang-Nguyen, D.T.; Dao, M.S. Multimodal Cheapfakes Detection by Utilizing Image Captioning for Global Context. In Proceedings of the 3rd ACM Workshop on Intelligent Cross-Data Analysis and Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 9–16.
44. Akgul, T.; Civelek, T.E.; Ugur, D.; Begen, A.C. COSMOS on Steroids: A Cheap Detector for Cheapfakes. In Proceedings of the 12th ACM Multimedia Systems Conference, Istanbul, Turkey, 28 September–1 October 2021; pp. 327–331.
45. Boididou, C.; Andreadou, K.; Papadopoulos, S.; Dang-Nguyen, D.T.; Boato, G.; Riegler, M.; Kompatsiaris, Y. Verifying multimedia use at mediaeval 2015. *MediaEval* **2015**, *3*, 7.