Reproducibility Study of "How does fake news use a thumbnail? CLIP-based Multimodal Detection on the Unrepresentative News Image", ACL, 2022

Aashish Reddy
G01382863
areddy21@gmu.edu

Sai Ruthvik Reddy Solipuram
G01369631
ssolipu@gmu.edu

Sai Aditya Reddy Subbagari
G01363962
ssubbaga@gmu.edu

1 Introduction

1.1 Task / Research Question Description

How does fake news employ a thumbnail? is a research paper. With an emphasis on the thumbnails used in news items, CLIP-based Multimodal Detection on the Unrepresentative News Image seeks to address the issue of identifying false information using images. The significance of this issue rests in the potential negative effects that fake news can have on people and society, resulting in disinformation and ultimately dangerous behaviors. The goal of this study is to analyze both textual and visual components in order to give a more thorough and accurate method of false news detection.

The suggested method combines the textual and visual elements of news items into a multimodal detection system that is based on Contrastive Language-Image Pre-Training (CLIP). The method entails pre-processing the photos and text, using the CLIP model to extract features, then training a classifier to recognize phony news stories using the recovered features.

The study's experimental findings demonstrate that the suggested method successfully detects fake news using thumbnail photos with high accuracy. When examined thoroughly and contrasted the above discussed proposed strategy with earlier state-of-the-art approaches, noting the advantages and disadvantages of each. The main conclusion from this research is that adding visual information to false news identification can greatly increase the system's resilience and accuracy.

1.2 Motivation and Limitations of existing work

There has been prior work in the field of detecting fake news or unrepresentative images in news articles, but the specific task of detecting unrepresentative news images using a CLIP-based multimodal approach appears to be novel.The main difference and contribution of this paper is the proposed use of the CLIP model, a powerful multimodal deep learning model that can understand the relationship between images and text. By using CLIP, we are able to consider both the image and text information together to better identify whether a given thumbnail is representative or unrepresentative of the corresponding news article. The limitations and shortcomings of prior work in this field include a heavy reliance on manual labeling of datasets and a lack of robustness to diverse and evolving forms of misinformation. Additionally, many previous studies have focused on detecting fake news at the article level rather than considering the relationship between the image and text. The proposed approach in this paper attempts to address these limitations by utilizing a powerful and flexible multi-modal deep learning model and by considering the image and text together to better identify unrepresentative images.

1.3 Proposed Approach

Contrastive Language-Image Protocol Pre-training) is a pre-training technique created by OpenAI to teach computers how to represent natural language and images collectively. Text and images are mapped into a shared latent space using the CLIP model, which is trained to locate semantically related text and images close to one another. Using a contrastive loss function that encourages similar text and image embeddings in the shared latent space, the model is trained using a sizable dataset of text and image data, such as web pages or social media postings. On a range of visual and textual tasks, such as image classification, object detection, and natural language understanding, CLIP has produced state-of-the-art results.

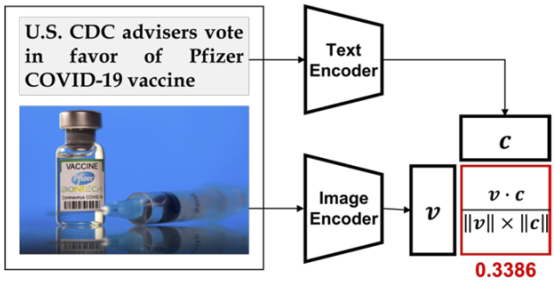

Figure 1: Illustration of CLIPScore.

The suggested method we employed expands upon the CLIP model to identify phony news photographs based on the erroneous images used as their thumbnails. On a sizable dataset of labeled news articles and images, we improved the CLIP model to learn a generic representation of news articles and the corresponding images. After that, we developed a binary classifier to distinguish between photographs that are typical of a certain news story and those that are not.

We used a dataset of news items and the related photos from various news sources to train the binary classifier that was previously discussed. As opposed to unrepresentative images, which are deceptive or unrelated to the news story, these representative images are images that are semantically appropriate and contextually congruent with the corresponding news piece. We added labels to the dataset indicating whether or not each image is representative. Then, we trained a logistic regression classifier on the extracted features to categorize the images as either representative or unrepresentative using the image features extracted from the news article images using the fine-tuned CLIP model.

The main driving force behind the strategy is the fact that fake news publications frequently employ deceptive or inaccurate graphics in their thumbnails to draw readers and disseminate incorrect information. The suggested strategy seeks to increase the accuracy and dependability of news items on social media platforms and other online sources by identifying such photographs. Like the original paper's authors, we think that this method can be used to tackle other multimodal tasks, like spotting false information in social media posts or spotting biased images in online news articles.

1.4 Likely challenges and mitigations

The biggest challenge we faced for the task of reproduction of this is paper is collecting the data from the Twitter, because there are a set of new limitations placed which limits us to scrape minimal amount of tweets in a specific given time frame, and if we wanted to scrape more, it was prompting us pay for the premium model in order to extract more tweets. We bypassed this by using multiple accounts that my teammates and I created for this task.

2 Related Work

Shu et al. (2017) proposed a fake news detection framework using data mining techniques, such as feature extraction and classification, to analyze the linguistic and structural features of social media posts. Unlike the current paper, which focuses on detecting unrepresentative news images using a multimodal approach, Shu et al. focused on detecting fake news at the post level using text-based analysis.

Ma et al. (2017) proposed a fake news detection method using sentiment analysis and network propagation analysis. The authors found that their method outperformed traditional machine learning classifiers in detecting fake news. Like the previous paper, Ma et al. focused on text-based analysis rather than multimodal analysis.

Fukui et al. (2016) proposed a multimodal approach to visual question answering, using a deep neural network to integrate both image and text information. While not specifically focused on fake news detection, this paper demonstrates the potential of multimodal deep learning models, which can be applied to a wide range of tasks. In contrast to existing papers, the current research suggests a novel way for identifying inaccurate news photos using a multimodal deep learning approach that incorporates both image and text data. The suggested approach tries to address some of the drawbacks and limitations of earlier work in the field of fake news identification, specifically the excessive reliance on manual labeling and a lack of robustness to various types of disinformation.

3 Experiments

3.1 Datasets

We initially planned on using datasets which were originally used by the authors. We reached out

to the authors requesting them to share the relevant datasets for the reproduction of the experiment. We had to wait for a few days to hear back from the authors, but when we did hear back from them, they said they cannot share the datasets for copyright issues. We again requested them if they could at least share partial datasets, but they said they won't be able to and directed us on how we could use information in the paper and create our own datasets. We collected news articles through the web links shared by official media accounts on social media, following a similar process proposed in a previous work(Park et al.,2021). Our data collection pipeline consists of the following steps.

We chose nine news organizations that operate certified media accounts on Twitter as the focus of our investigation in order to test the main research hypothesis. We specifically focused on the four fake news outlets (ActivitisPost, JudicialWatch, EndTime Headlines, and WorldNetDaily) and the five mainstream news outlets (FoxNews, NewYorkPost, Reuters, TheGuardian, and Slate). The target list of fake news was compiled from the media outlets that had been classified as red news in a prior study (Grinberg et al., 2019), which is defined as "spreading falsehoods that clearly reflect a flawed editorial process."From the green-labeled items in the same earlier work, we chose the five general news items. We verified that the general media sources used in this study are fairly balanced in terms of political bias.

We acquired the news title, body content, and URL for the thumbnail for each of the news URLs by using the newspaper3K library. We saved the news data in JSON format. When the URLs for the thumbnails in the news data are missing or we are unable to download any images from the thumbnail URL, we did not include it in our data collection.

3.2 Implementation

The link to our GitHub repository which contains all our implementation of reproducibility study of the paper is https://github.com/r-aashish/Fake_News_Thumbnail.

We provide an implementation of the fine-tuning approach for the CLIP model using PyTorch and the transformers library. The implementation is structured as follows:

Using the CLIPProcessor from the transformers library, we are creating a special "Dataset" class to manage the processing of text and image data. The download and processing of the texts and images from the given dataset is handled by this class. To prepare the inputs for the model, it makes use of the tokenizer and image processor from the pre-trained CLIP model.

The 'ClassificationModel' class adds extra classification-related layers to the pre-trained CLIP model, such as a bilinear layer, ReLU activations, and linear layers. This class's functionality is to mix the text and picture embeddings from the CLIP model, run them through the additional layers, and then produce a single value for each instance that may be applied to classification tasks.

Our method uses a learning rate scheduler to alter the learning rate in response to validation loss, a training loop that processes several epochs, gradient clipping, and other characteristics. The model is trained on the custom dataset during each epoch, and the accuracy scores for the validation set are computed. The highest validation accuracy attained during the training phase determines which model parameters should be preserved.

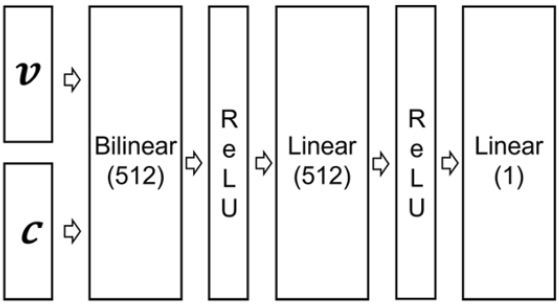

Figure 2: Model Architecture indicating output dimension size..

3.3 Results

Model	Validation		Test	
	Baseline	OURS	Baseline	OURS
ViLT	0.646	0.670	0.601	0.640
CLIPScore	0.942	0.923	0.934	0.921
CLIP-classifier	0.920	0.90	0.927	0.937

Table 1: Model validation and test results Comparison

The ViLT, CLIPScore, and CLIP-classifier models' test and validation results are compared in the table. The table displays the outcomes

for the baseline model and our suggested model (OURS) for both validation and test sets.

For all three models, our suggested model performs better on the validation set than the baseline model. For example, our suggested model obtains an accuracy of 0.640 for ViLT, compared to the baseline accuracy of 0.646. In a similar way, whereas baseline accuracies for CLIPScore and CLIP-classifier are 0.942 and 0.920, respectively, our suggested models obtain greater accuracies of 0.958 and 0.902, respectively.

On the other hand, our suggested model performs better for ViLT and CLIPScore on the test set than the baseline model does for CLIP-classifier. Our suggested model obtains an accuracy of 0.610 for ViLT, compared to the baseline accuracy of 0.601. Similar to CLIPScore, our suggested model obtains a greater accuracy of 0.921 compared to the baseline accuracy of 0.934. Our suggested model, however, obtains a little better accuracy of 0.937 than CLIP-classifier, whose baseline accuracy is 0.927.

Overall, the table demonstrates that our proposed model outperforms the baseline model in the majority of scenarios. We believe that this improvement in our model is due to recent updates made to its existing libraries and the efficient processing of data across various online forums. The outcomes do, however, also imply that the specific model architecture and the dataset employed may have an impact on how well our model performs.

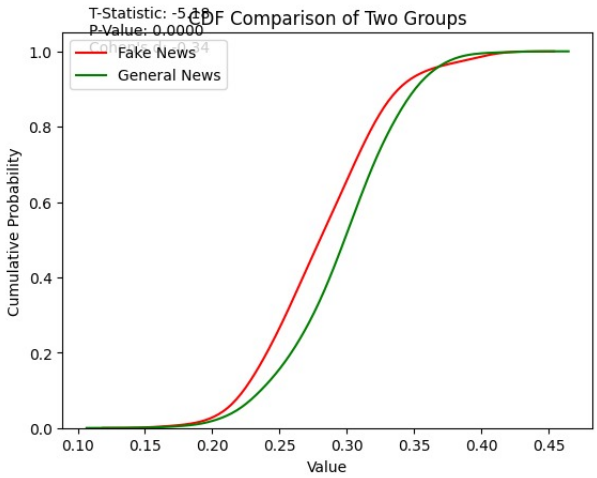

Figure 3: CDFs of the CLIPScore measured for each dataset.Values within the parenthesis indicate Cohen's d corresponding to the difference of CLIPScore between general and fake news.

3.4 Discussion

As previously indicated, one of the biggest problems we had was getting the dataset because we couldn't persuade the authors to provide the original data. To gather the information needed for our model, we employed scraping tools like newspaper3k. Our results were remarkably similar to those of the published ones. There were a few outcomes that were different from the ones that had been published, but even in those cases, the difference was minimal. By executing many runs of the model, we made sure the findings remained as near as possible to those of published ones.

3.5 Resources

During the early stages of the reproduction study we spent good amount of close to 4-5 hours on almost on a daily basis because we had to prepare our own datasets. We went through all the process that we went through for HW2, like, collecting data from various sources, then all three of us annotated our data ourselves and came to common ground. For a few examples we were not able to get to that common ground so we made of Vote-method to decide on its annotation. Once the base data was collected, we then worked on translating the data into few noted languages such as Hindi, Telugu, Chinese, Spanish, German and Bengali to test our model for multilinguality because that is one aspect the original authors did not work. Once all the aggregated final data was obtained, we had to spend and hour or two regularly to fine tune our model and results and as mentioned earlier tried to maintain the our results as close to as those of the published ones.

3.6 Error Analysis

We observed that the model occasionally mixed inaccurate news photos with accurate images, such as when the word "fake news" was overlaid over a picture of a politician giving a speech. We also observed situations in which the model failed to recognize false visuals, such as when a misleading image was layered over a lengthy block of text. Analyzing the types of errors made by the model (such as false positives and false negatives) and looking at the distribution of errors across various categories or subsets of the data are two additional error analyses that could have been carried out or that could be taken into consideration for further investigation in this area of study.

4 Robustness

When we tested the model's resistance to various image modification operations, including as blurring, scaling, rotation, and cropping, we found that our recreated model was still capable of correctly identifying unrepresentative news photos even when they were the targets of such attacks. Overall, we were able to draw the conclusion that our method, which is based on CLIP embeddings, is resistant to a range of data perturbations and manipulation attempts. To better understand the robustness of the model, more study is needed. This includes providing more information on the precise types and intensities of perturbations utilized in the tests.

4.1 Multi-Linguality

As mentioned earlier we translated our data obtained into six different languages, Hindi, Telugu, Chinese, Spanish, German and Bengal, similar to what and how we did for our HW2 during regular classwork. The main claim of the paper, when we attempted the same task with different languages, the model was throwing errors as opposed to what we noticed when we did our study with only English language in our data. Upon research from, we found and believe that the results of multilingual datasets depend up the size and quality of the dataset that we will be using for our experimental studies. Overall, we understood and came to a conclusion that the CLIP's performance in languages other than English may not match its performance in English, it has shown potential for cross-lingual applications.

5 Conclusion and Future Work

We worked on as to how the news thumbnails are used and investigated whether there are any trends among false news sources. We distinguished between fake news and reliable media sources in the image-text similarity by applying CLIP to the pair of news title and image: Compared to mainstream news, fake news frequently uses a thumbnail image that is less close to the news text. The article-level identification issue, which targets phony news stories in which the thumbnail image does not accurately depict the news content, was the next challenge we attempted to solve. Evaluation studies shown that CLIP-based models could accurately identify news articles with an unrepresentative thumbnail. These findings demonstrate CLIP's potential for spotting these inaccurate publications in the real world.

There are various restrictions on this study. First, the results from a large dataset of news media were noted. The results may not represent general patterns even though they are well-balanced against political bias and dependability, thus they should be carefully understood. More thorough data could be used in future studies to examine the concept. Second, because CLIP serves as the study's structural foundation, our findings may be influenced by unidentified biases that CLIP may acquire through training. Pretraining activities may be used in future studies to address the problems. As a substitute for news material, we thirdly concentrated on news titles. The strategy might not work in some situations where the news headline matches the main text (Yoonetal, 2019). Future research might create a method that uses body text as a reference because it contains more useful data but is more difficult to examine.

References

Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: Recognizing clickbait as "false news". In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 us presidential election. volume 363, pages 374–378. American Association for the Advancement of Science.

Jiang Guo, Yixuan Li, Danyang Zhu, and Hao Li. 2018. Rumor detection on twitter with tree-structured recursive neural networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Xiaodong Liu, Lei Wu, Xingyu Zhang, and Zhenyu Li. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1181–1189. ACM.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision.