I-KAHAN: Image-Enhanced Knowledge-Aware Hierarchical Attention Network for Multi-modal Fake News Detection

Øystein L. Nilsen[1], Pelin Mişe[2], Ahmet Yıldız[2], Eniafe Festus Ayetiran[3,4][0000−0002−6816−2781], and Özlem Özgöbek[3][0000−0003−2612−2009]

[1] Sopra Steria, Oslo, Norway
oystein.lnilsen@soprasteria.com

[2] MEF University, Department of Computer Engineering, Istanbul, Turkey
{misepe, yildizah}@mef.edu.tr

[3] Norwegian University of Science and Technology (NTNU), Department of Computer Science, Trondheim, Norway
{ozlem.ozgobek, eniafe.ayetiran}@ntnu.no

[4] Achievers University, Department of Computer Science, Owo, Nigeria
eniafe.ayetiran@achievers.edu.ng

Abstract. In the quest to combat the proliferation of fake news, accurate detection of fabricated news content has become increasingly desirable. While existing methodologies leverage a variety of news attributes, such as text content and social media comments, few incorporate diverse features from different modalities like images. In this paper, Image-Enhanced Knowledge-Aware Hierarchical Attention Network (I-KAHAN) architecture is proposed as an enhancement to the existing KAHAN architecture. The I-KAHAN architecture utilizes a wide variety of attributes including news content, user comments, external knowledge, and temporal information which are inherited from the KAHAN architecture, and extends it by integrating image-based information as an additional feature. This work contributes to refining and expanding fake news detection methodologies by embracing a more comprehensive range of features and modalities, and offers valuable insights into the effectiveness of various methods for the numerical representation of images, feature aggregation and dimensionality reduction. Experiments conducted on two real-world datasets, PolitiFact and GossipCop, assessing the performance of the I-KAHAN architecture, demonstrated approximately 3% improvement in accuracy over the KAHAN architecture, highlighting the potential benefits of incorporating diverse features and modalities for enhanced fake news detection performance.

Keywords: Fake News Detection · Deep Learning · Multi-Modality.

1 Introduction

Rapid advancements in digital technology gives people easier access to news via websites or social media. Even though this may be advantageous, it also brings challenges like the rapid spread of fake news which may have serious real world consequences. A sequence of significant occurrences, notably highlighted by the 2016 US presidential election, have intensified the gravity of the challenge posed by disinformation [1]. A few years after, during the COVID-19 pandemic, the director-general of the World Health Organization (WHO) characterized the extensive proliferation of false information throughout the pandemic not just as a battle against a disease, but also as an "infodemic", which lead to active campaigns by WHO to fight against it [3], [4]. During the Covid-19 pandemic, the impact of fake news on public health has shown the extent of possible serious consequences [2]. More recently, the Russian incursion into Ukraine further compounded the fake news problem. The various instances of fabricated videos and images emerging from the conflict between Russia and Ukraine, serve as concrete examples of how disinformation is being exploited as a deliberate tool in the realm of warfare, with the intention of manipulating the prevailing narrative [5].

Given the potential global and local implications of fake news, it is evident that the swift and effective detection of such news is of utmost importance. The concept of fake news and automated fake news detection systems has been increasingly in the focus of researchers in the past years. However, to the best of our knowledge, these systems are still not good enough to be widely deployed in real world applications.

In the context of news, images are powerful additions to text. There have been attempts to understand the connection between the textual content of news articles and the images accompanying them [14]. Usually, images are used to enhance the meaning or descriptions of the events mentioned in the news article. Images may also amplitude the feelings of readers, making the news articles more appealing or easier to relate to. Fake news articles are known to include certain features such as emotions more than non-fake articles [15]. These features (e.g. more powerful emotions) may be supported by images. Also the descriptive addition to news articles (e.g. a photo describing a recent event or representing a person) may be different in fake news articles. Therefore, in addition to focusing only on the textual content for automated fake news detection, incorporating the information from images may give us more clues about the seriousness or correctness of an article. In most cases, multi-modal fake news detection method rely on this possible connection.

In this study, multi-modal fake news detection architecture, Image-Enhanced Knowledge-Aware Hierarchical Attention Network (I-KAHAN), is proposed as an extension of an existing text base fake news detection system called Knowledge-Aware Hierarchical Attention Network (KAHAN) [6]. In addition to news content and social media comments, I-KAHAN incorporates news images as an additional feature. This multi-modal approach, along with the use of deep neural networks, aims to enhance the performance of the fake news classifier. In this study, the various configurations of I-KAHAN, using different methods of image embedding, dimensionality reduction and feature fusion methods, are thoroughly investigated. A novel dimensionality reduction method, Image-based Hierarchical Attention Network (IHAN) which incorporates attention mechanisms for dimensionality reduction is proposed. The effectiveness of various configurations of I-KAHAN is evaluated using two real-world datasets, namely PolitiFact and GossipCop.

The paper is organized as follows: Related work is presented in Section 2. The methodology is explained and I-KAHAN architecture is presented in Section 3. Experimental details and results are presented in Section 4, along with the discussions. Finally Section 5 gives conclusions and directions for future work.

2 Related Work

Numerous studies have focused on fake news detection using both uni-modal and multi-modal approaches. Uni-modal approaches mostly use the textual content of news articles as the single data source, whereas in the multi-modal approaches images are usually the most common accompanying data form.

To better capture relevant information which is useful for detecting fake news, KAHAN [6] specifically introduces user comments in addition to the main news text into a time-based sub-event division attention-based model. This approach provides pattern from user comments, augmenting the detection process by providing an understanding of user interaction with the news content. The architecture has four main components which are user comment encoder that models the user comments, news content encoder that models the text of the news, an external knowledge attention mechanism that makes use of additional data from a knowledge graph and the fake news classifier. KAHAN highlights the importance of the external knowledge on classification efficacy with improved performance over previous works with similar ideas.

The Knowledge-Aligned News model (KAN) [7] is with the introduction of external knowledge through knowledge graphs to better capture the relationship among news entities, textual contents and social context of the news. The model uses additional data to its analysis by matching entities in the news content with entries in the knowledge graph. However, research has shown that the effectiveness of KAN depends on how good and complete the knowledge graph being used is. This emphasizes the value of auxiliary knowledge while simultaneously highlighting the need for caution due to potential flaws in the accuracy and comprehensiveness of the underlying graph.

The shift from uni-modal to multi-modal techniques allows for the use of a wider contextual landscape than just textual material. FakeMine [10] is a multi-modal approach which focuses on fake news detection in social media with particular emphasis on the network structure of social media posts. FakeMine effectively combines several modalities, such as text and image, by using a graph neural network to examine social media post networks. In addition, it uses BERT [17] for textual content representation, and VGG19 for image characteristics. FakeMine exhibits improvements with the addition of an LSTM classifier that was enhanced utilizing the Chimp optimization technique.

The Event Adversarial Neural Network (EANN) [11] is another approach to detect fake news. It extracts features that are independent of the event, which helps to detect fake news related to recently occurring news. EANN has three primary components which are multi-modal feature extractor, a fake news detector and an event discriminator. In order to accurately detect fake news, the multi-modal feature extractor works collaboratively with the fake news detector to acquire various representations of the textual and visual attributes.

Sentiment-Aware Multi-modal Embedding (SAME) [12], which maintains semantic relevance across modalities, also adds user sentiment into the detection of fake news. One of the most recent works in deep learning-based approach to multi-modal fake news detection is the work of [13] which in order to remove the heterogeneity and semantic gap inherent in multi-modal understanding unifies the modalities using an inter-modal attention-based BiLSTM-CNN architecture.

In addition, (dEFEND) [8] posits that explainable detection of fake news is an important factor missing in similar studies. Their method uses a sentence-comment co-attention sub-networks to jointly capture explainable check-worthy sentences and user comments. [22] presents a comprehensive survey of diverse approaches to fake news detection stressing their characteristics, techniques and datasets.

3 Methodology

This study presents an innovative architecture, Image Enhanced Knowledge Aware Hierarchical Attention Network (I-KAHAN), which is an extension of the KAHAN architecture. The I-KAHAN architecture improves the KAHAN model on the incorporation of multi-modality by introducing news images as an additional feature and the enhancement of classifier performance by employing deep neural networks. Figure 1 provides an overview of the proposed I-KAHAN architecture. The I-KAHAN architecture leverages three primary data sources: news content, user comments, and images, to classify news articles as either fake or real. The I-KAHAN architecture employs the same approach for external knowledge extraction and textual content (both news content and comments) processes as the original KAHAN architecture. For obtaining a proper form from the user comments and contents of the news, pre-trained models GloVe [19] and Wikipedia2vec [20] was used for text embeddings. Two GloVe models, *glove-wiki-gigaword-100* and *glove-twitter-100* were used for news content and user comments respectively. The *nwiki-20180420-100d* through Wikipedia2vec was used for entity and entity claim embeddings. For the images, three main processes which are image embeddings, dimensionality reduction and feature fusion were done to find the best configurations. At the end of these, a fused feature vector is obtained. The details of each component of the I-KAHAN architecture is explained in the following sub-sections.

When introducing news images as an additional feature, several challenges arise, which can be classified into three categories: Embedding, Dimensional-

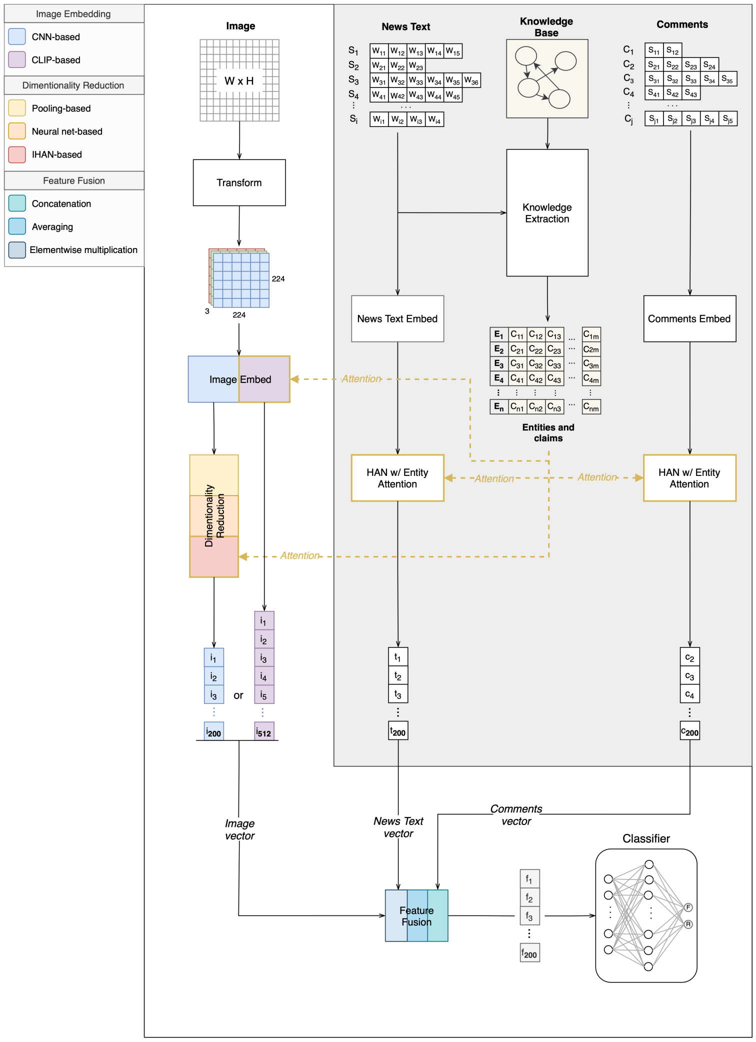


Image Embedding
CNN-based
CLIP-based
Dimentionality Reduction
Pooling-based
Neural net-based
IHAN-based
Feature Fusion
Concatenation
Averaging
Elementwise multiplication
Image
W x H
Transform
224
3
224
Image Embed
Dimentionality Reduction
i_1 i_2 i_3 ⋮ i_{200}
or
i_1 i_2 i_3 i_4 i_5 ⋮ i_{512}
News Text
S_1 W_{11} W_{12} W_{13} W_{14} W_{15}
S_2 W_{21} W_{22} W_{23}
S_3 W_{31} W_{32} W_{33} W_{34} W_{35} W_{36}
S_4 W_{41} W_{42} W_{43} W_{44} W_{45}
S_i W_{i1} W_{i2} W_{i3} W_{i4}
Knowledge Base
Knowledge Extraction
Comments
C_1 S_{11} S_{12}
C_2 S_{21} S_{22} S_{23} S_{24}
C_3 S_{31} S_{32} S_{33} S_{34} S_{35}
C_4 S_{41} S_{42} S_{43}
C_j S_{j1} S_{j2} S_{j3} S_{j4} S_{j5}
News Text Embed
Comments Embed
E_1 C_{11} C_{12} C_{13} … C_{1m}
E_2 C_{21} C_{22} C_{23} … C2m
E_3 C_{31} C_{32} C_{33} … C_{3m}
E_4 C_{41} C_{42} C_{43} … C_{4m}
E_n C_{n1} C_{n2} C_{n3} … C_{nm}
Entities and claims
Attention
Attention
Attention
Attention
HAN w/ Entity Attention
HAN w/ Entity Attention
t_1 t_2 t_3 ⋮ t_{200}
c_2 c_3 c_4 ⋮ c_{200}
Image vector
News Text vector
Comments vector
Feature Fusion
f_1 f_2 f_3 ⋮ f_{200}
Classifier
F
R

Fig. 1. Overview of the I-KAHAN architecture

ity Reduction, and Fusion. Due to multi-modality, these challenges have been addressed to improve the fake news detection system's efficiency.

3.1 Image Embedding Methods

Images cannot be comprehended by machines in the same way that people do. Images must therefore be expressed numerically to be understood by machines. There are numerous techniques that can be used for this. In this study, images have been converted into numeric arrays using CLIP models and pre-trained Convolutional Neural Networks (CNNs) like VGG19 [21] and ResNet-50 [18] .

Deep CNN-based Embeddings VGG19 and ResNet-50, two well-known deep convolutional neural network (CNN) architectures, were used in this study. These architectures were chosen because they have a track record of success in object recognition tasks and are frequently used in fields like fake news detection. The final convolutional layer of each architecture yields the resulting feature vector. An input image must be transformed before being sent through the VGG19 architecture for embedding in the feature extraction process. By skipping the categorization stage and concentrating simply on capturing distinctive traits, this results in a feature vector with a size of 25088. The ResNet-50-based feature extraction adopts a similar methodology. The image is altered and then embedded using the ResNet-50 architecture, producing a feature vector with a size of 100352.

CLIP Embeddings Contrastive Language-Image Pretraining (CLIP) [16] is a recent technique that jointly learns image and text representations in a shared embedding space. With its unique ability to encrypt both text and images in a single common vector space, CLIP creates brief but detailed embeddings that closely resemble textual features. This study experiments with two different CLIP implementations. The altered image is directly encoded using CLIP in the first method, producing an image vector that is located in the common vector space. In the second method, individual entities and claims are embedded using CLIP's text encoder. An attention mechanism with multiple heads is then applied to the visual vector.

3.2 Dimensionality Reduction Methods

When numeric vectors have been created from the images using the pre-trained CNNs and CLIP approaches, the high-dimensionality of the resulting vectors makes it difficult to train models. Dimensionality reduction is therefore required for these produced vectors so that they have the same weighting in the classifier as the more compact text embeddings. In this study, neural network-based methods and pooling-based techniques are investigated to minimize the high-dimensional vectors acquired from image embedding. In addition, we propose a novel Image-based Hierarchical Attention Network (IHAN) approach for dimensionality reduction and compare it to aforementioned approaches. The results are given in Section 4.3.

Neural Network-based Dimensionality Reduction One of the methods that was used in this study's dimensionality reduction procedure is the Neural Network based approach. Utilizing the image vectors derived from CNN-based image embeddings as input to these neural networks is the method. These networks then produce 200-dimensional reduced image vectors with 200 dimensions apiece.

Pooling-based Dimensionality Reduction Pooling techniques, such as max pooling and average pooling, are considered as alternative methods for reducing the dimensionality of image embeddings. These techniques aggregate the information within the feature maps by applying a pooling operation, which can effectively reduce the size of the embeddings while preserving relevant information.

Image-based Hierarchical Attention Network (IHAN) The Image-based Hierarchical Attention Network (IHAN) introduces a novel approach to image processing, adapting attention mechanisms typically used for dimensionality reduction in textual data to the realm of images. Essentially, IHAN is an advanced multi-level pooling technique that employs the core principles of the Hierarchical Attention Network (HAN) in the field of image processing. IHAN is unique due to its three primary characteristics: Hierarchical Structure, Multi-Level Pooling, and Entity-Attention Inspired Layer. Recognizing that images, much like written text, have a hierarchical structure, IHAN applies a similar layered approach to image processing. This involves using a multi-level pooling mechanism that mirrors the hierarchical attention employed in HAN for textual data. This method enables IHAN to incrementally extract information from smaller segments, akin to "words", to larger ones, similar to "sentences", within an image. Further enhancing its capabilities, IHAN incorporates an Entity-Attention Inspired Layer to improve its image embeddings. This layer prioritizes contextually relevant visual information, improving IHAN's ability to effectively comprehend and interpret the visual content. This study investigates the application of IHAN in the I-KAHAN architecture, both with and without the entity attention mechanism, thereby highlighting its versatility and adaptability in handling varying image processing requirements.

3.3 Feature Fusion Methods

The final challenge is fusing the image features with other features effectively. Three distinct techniques for feature fusion were investigated: concatenation, element-wise multiplication, and averaging.

3.4 Fake News Classifier

The final step in I-KAHAN involves feeding the fused feature vectors to a classifier. This classifier is a feed-forward neural network, similar to the one in the KAHAN model. It performs binary classification, outputting a probability distribution over two classes: "fake" or "real." Two alternative classifier architectures have been implemented, one with a single hidden layer (shallow classifier), as in KAHAN, and another with two hidden layers (deep classifier). Within its architecture, I-KAHAN introduces the innovative IHAN method, which leverages attention mechanisms in a creative way and demonstrates impressive performance.

4 Experiments and Results

4.1 Datasets

PolitiFact and GossipCop, two separate datasets are both integral to the FakeNewsNet dataset [23]. It was observed that in a subset of the PolitiFact dataset, real news images frequently feature politicians and reputable news organizations, while fake news images frequently overcharge and stir up controversy. The distinction between real and fake news in the GossipCop dataset is not as clear-cut as it is in the PolitiFact dataset, despite some visual clues that suggest it. Some modifications of the original strategy were used to gather the news content in order to assure the highest quality of the data, strengthening the dataset's integrity and the validity of the conclusions reached from it. Some of the news in the datasets were removed after data cleaning processes due to unrelated images with news such as logos and advertisements. The total number of news items in the final datasets are presented in Table 1.

Table 1. Statistical information derived from the PolitiFact and GossipCop datasets subsequent to the completion of data cleaning.

	PolitiFact	GossipCop
Real News	219	1564
Fake News	172	1779
Total News	**391**	**3343**

4.2 Experimental Setup

All the experiments in this study has been run on IDUN HPC cluster at the Norwegian University of Science and Technology[5]. The array jobs were dispatched with one CPU core each and 20 GB of memory.

As mentioned in Section 3, the I-KAHAN architecture uses two *pre-trained embedding models*, GloVe[6] and Wikipedia2vec[7] for text and entity embeddings respectively.

The *hyperparameters* for the I-KAHAN model are presented in the Table 2. They have been kept consistent across all experiments to ensure an equitable comparison among different configurations.

The *evaluation metrics* used for the experiments are accuracy, precision, recall and F1 scores, with 3-fold cross validation.

The implementation details and code of I-KAHAN can be found on Github[8].

[5] https://www.hpc.ntnu.no/idun/

[6] https://nlp.stanford.edu/projects/glove/

[7] https://wikipedia2vec.github.io/wikipedia2vec/

[8] https://github.com/oysteinlondal/I-KAHAN

Table 2. Hyperparameters utilized in all I-KAHAN model experiments.

Hyperparameter	**Value**
Epochs	65
Batch Size	16
Learning Rate	$5 * 10^{-5}$
Number of Seeds	3
Number of Folds	3
Hidden Size	100
Weight Decay	$1 * 10^{-4}$
Dropout	0.3

4.3 Performance Comparison of Methods

One of the key points of the study comprised a thorough assessment of the effectiveness of various methods and techniques integrated into the I-KAHAN architecture, followed by the determination of the most effective combination. For the complexity of both datasets (PolitiFact and GossipCop), the inquiry involved experimenting with various strategies and approaches under each method. Used techniques and their performance comparisons in both datasets can be seen in Figures 2, 3 and 4.

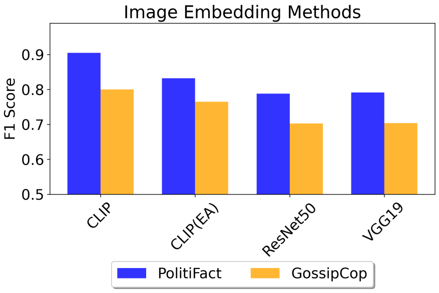
In case of image embedding methods, it was observed that in both datasets CLIP outperformed the other methods. Additionally, the novel dimensionality reduction method we proposed, IHAN, showed the best results without the use of attention on both datasets, matching the performance levels of the best apprach, namely pooling. In case of feature fusion method comparison, concatenation outperformed the other methods.

Throughout the process, the unique properties of each dataset led to the discovery of combinations that maximized performance for both. In this study, different combinations of these methods were experimented and compared to find out the best performing combination. As explained in Section 3.4, the experiments were run with two different type of classifiers: Deep and shallow. The comparisons of these different classifiers and combinations are presented in Section 4.4. Results show that specific combinations of these methods showed the potential to considerably improve fake news detection overall, outperforming the baseline set by the KAHAN model.

4.4 Fake News Detection Performance Results

In this section, results from two different classifiers which are shallow and deep classifiers and several different combinations of methods are presented.

The first objective was to determine whether the deep classifier might improve these configurations' performance. The shallow classifier typically matches or outperforms the deeper models, particularly in the GossipCop dataset as it can be seen Figure 6. For the PolitiFact dataset deep classifier seems to outperform the shallow classifier for some of the configurations as shown in Figure



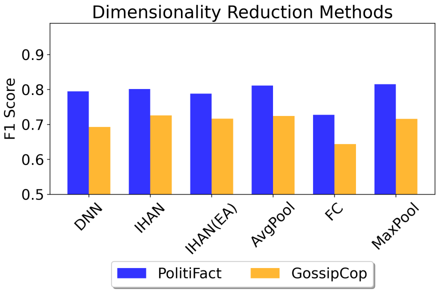


Fig. 2. Comparison of Image Embedding Methods

Fig. 3. Comparison of Dimensionality Reduction Methods

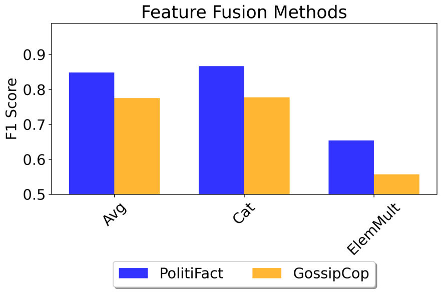


Fig. 4. Comparison of Feature Fusion Methods

5. The performance of the deep and shallow classifiers for different configurations considering embeddings, dimensionality reduction and feature fusion are shown in Tables 3 and 4. For both datasets, a shallow classifier performs the best, however with different configurations. The *CLIP-Cat* configuration performs the best when using the PolitiFact dataset, whereas the *ResNet50-DNN-Avg* configuration performs the best when using the GossipCop dataset.

Confusion matrices are presented for each of the top-performing combinations to provide a deeper understanding of their performance. During k-fold cross-validation, each matrix captures the classification performance of a certain fold. Figure 7 represents the *CLIP-Cat* confusion matrix for PolitiFact, indicating a low number of false negatives and positives. It's interesting to note that true positives outnumber true negatives by 43% to 52%. In Figure 8, the confusion matrix for GossipCop shows that there are more true negatives than positives and slightly higher rates of false negatives and positives. This trend is reversed. Basically, these confusion matrices propose that *CLIP-Cat* inclines towards ordering news all the more frequently as real, while *ResNet50-DNN-Avg* will in general group news all the more frequently as fake. However, a significant point to consider is the data imbalances favoring real news in Politifact and fake news in GossipCop. This discrepancy is likely the most plausible explanation for the differing trends observed in the two models, potentially skewing the results and influencing the classification tendencies of each model.

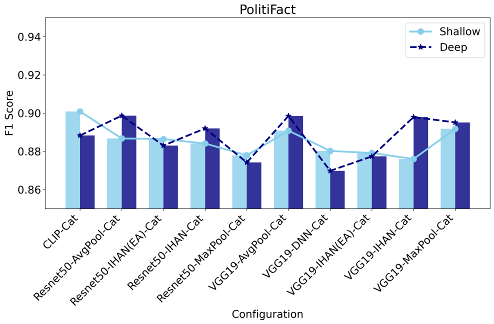


Fig. 5. Performance comparison of the Shallow and Deep classifier for PolitiFact

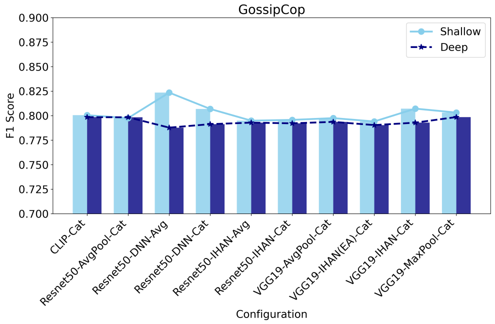


Fig. 6. Performance comparison of the Shallow and Deep classifier for GossipCop

Table 3. Detailed comparison of the shallow classifier on the PolitiFact and GossipCop datasets.

	PolitiFact				GossipCop			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
CLIP-Cat	**0.9020**	**0.9059**	**0.8983**	**0.901**	**0.8011**	**0.8020**	**0.8001**	**0.800**
ResNet50-AvgPool-Cat	0.8875	0.8873	0.8863	0.887	0.7978	0.7973	0.798	0.798
ResNet50-IHAN-Cat	0.8850	0.8859	0.8836	0.884	0.7959	0.7965	0.7968	0.796
VGG19-AvgPool-Cat	0.8918	0.8922	0.8888	0.891	0.7977	0.7978	0.7984	0.797
VGG19-IHAN(EA)-Cat	0.8799	0.8798	0.8793	0.879	0.7944	0.7963	0.7954	0.794
VGG19-IHAN-Cat	0.8773	0.8789	0.87340	0.876	0.8073	0.8084	0.809	0.8071
VGG19-MaxPool-Cat	0.8926	0.8924	0.8911	0.892	0.8034	0.8041	0.8044	0.8031

Analyzing the loss per epoch graph provides valuable insights into the performance of the CLIP-Cat model. The graph displays a counter-intuitive phenomenon where the validation loss starts lower than the training loss. This is primarily due to the implementation of regularization and dropout techniques on the classifier. These techniques, which include modifying the algorithm and randomly ignoring select neurons during training, help reduce weight sensitivity and prevent overfitting. In the case of the PolitiFact dataset, the graph shows the training loss nearing zero, indicating overfitting. However, the model miti-

Table 4. Detailed comparison of the deep classifier on the PolitiFact and GossipCop datasets.

	PolitiFact				GossipCop			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
CLIP-Cat	0.8892	0.8911	0.8874	0.8883	0.7987	0.7986	0.7991	0.7984
ResNet50-AvgPool-Cat	0.8995	0.8988	0.8980	0.8987	0.7984	0.7999	0.8005	0.7983
ResNet50-IHAN-Cat	0.8926	0.8922	0.8923	0.8920	0.7924	0.7937	0.7935	0.7921
VGG19-AvgPool-Cat	0.8994	0.8997	0.8966	0.8985	0.7938	0.7938	0.7948	0.7937
VGG19-IHAN(EA)-Cat	0.8781	0.8771	0.8777	0.8773	0.7904	0.7926	0.7930	0.7904
VGG19-IHAN-Cat	0.8986	0.8976	0.8983	0.8979	0.7930	0.7934	0.7942	0.7928
VGG19-MaxPool-Cat	0.8960	0.8968	0.8940	0.8951	0.7987	0.7995	0.8002	0.7986

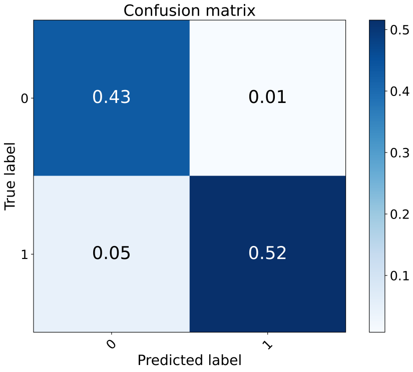


Fig. 7. Confusion matrix for CLIP-Cat on the PolitiFact

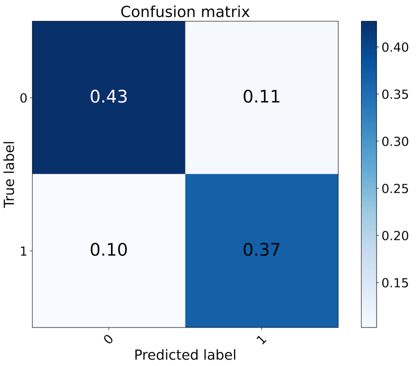


Fig. 8. Confusion matrix for ResNet50-DNN-Avg on the GossipCop

gates this through the use of a learning rate scheduler which adjusts the learning rate during training. This is evidenced by the plateauing of the graphs. Similar observations can be made with the GossipCop dataset, where the model appears less prone to overfitting, retaining a training loss slightly above zero. Here, the learning rate scheduler effectively maintains the validation loss at a steady level after the initial epochs.

In addition to these techniques, weight decay regularization was also employed to further minimize overfitting by reducing weights and complexity. The ReduceLROnPlateau strategy, which decreases the learning rate when the model's improvement plateaus, was implemented for added protection against overfitting. This strategy aids the model in converging to a global minimum and optimizing performance on unseen data.

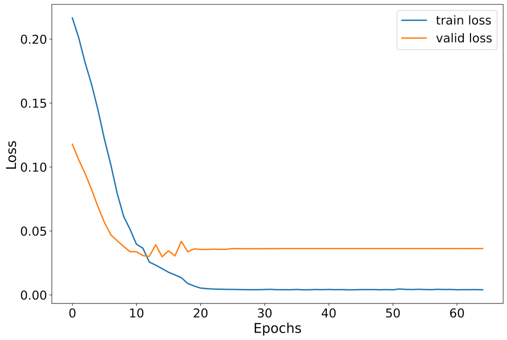


Fig. 9. Loss per epoch for CLIP-Cat on the PolitiFact

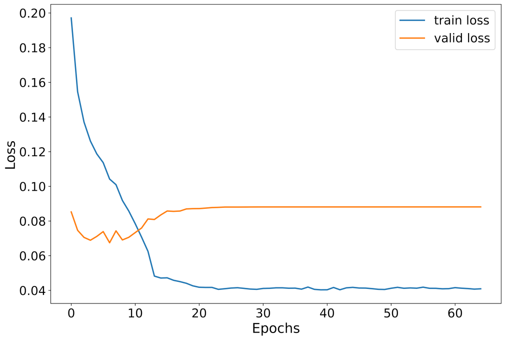


Fig. 10. Loss per epoch for CLIP-Cat on the GossipCop dataset

4.5 Comparison of I-KAHAN and KAHAN

To ensure the reliability of our results, a seeding model was employed to maintain consistency in scores across multiple runs, enhancing the robustness of our evaluation process. This process encompassed three rounds of three-fold validation using varied seeds. The scores referenced in this paper are the result of an averaging process conducted across these different folds and seeds, thus providing a more accurate and reliable representation of our findings.

5 Conclusion

In this study, we proposed a novel multi-modal fake news detection framework called the Image-enhanced Knowledge-Aware Hierarchical Attention Network (I-KAHAN) which was built on KAHAN, a prior architecture. The goal of this multi-modal approach is to investigate the effect of images when used along

Table 5. Comparison of the KAHAN and I-KAHAN Architectures

	PolitiFact				GossipCop			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
KAHAN	0.8756	0.8762	0.8732	0.8745	0.7894	0.7904	0.7905	0.7892
I-KAHAN	**0.9020**	**0.9059**	**0.8983**	**0.901**	**0.8011**	**0.8020**	**0.8001**	**0.8005**

with text in the detection of fake news. In addition, we proposed a novel dimensionality reduction method, Image-based Hierarchical Attention Network (IHAN) which includes an attention mechanism. The results show that different I-KAHAN configurations performed better on different datasets, with consistently improved outcomes shown in all cases on the PolitiFact dataset. The number of news items in the datasets and the quality of the news images were recognized as potential causes of this discrepancy, pointing to the importance of future data collection processes. It was discovered that the I-KAHAN design outperformed the KAHAN architecture in both datasets, with the PolitiFact dataset experiencing approximately 3% accuracy boost and the GossipCop dataset experiencing approximately 1% increase. This suggests that images do, albeit in a little way, help in multi-modal fake news detection. Our understanding of how various I-KAHAN setups operate on diverse datasets and how results may vary correspondingly might be further improved by future study.

6 Acknowledgement

This work was carried out during the tenure of the fourth author as an ERCIM “Alain Bensoussan” fellow.

References

1. Allcott, H., Gentzkow, M.,: Social Media and Fake News in the 2016 Election. Journal of Economic Perspectives, vol. 31, no. 2, pp. 211–236, May 2017, issn: 0895-3309. https://doi.org/10.1257/jep.31.2.211.
2. Rocha, Y., M., de Moura, G. A., Desidério, G. A., de Oliveira, C. H., Lourenço, F. D., de Figueiredo Nicolete,L. D.,: The impact of fake news on social media and its influence on health during the COVID-19 pandemic: A systematic review. Journal of Public Health, Oct. 2021, issn: 1613-2238. doi: 10.1007/s10389-021- 01658-z. https://doi.org/10.1007/s10389-021-01658-z
3. World Health Organization. (n.d.). Infodemic. Www.who.int. https://www.who.int/health-topics/infodemic
4. World Health Organization. (2020, September 23). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. World Health Organization. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

5. Shin, Y., Sojdehei, Y., Zheng, L., Blanchard, B.: Content-Based Unsupervised Fake News Detection on Ukraine-Russia War. SMU Data Science Review vol. 7, no. 1, Article 3. (2023).
6. Tseng, Y., Yang, H., Wang, Y., Peng, W.: KAHAN: Knowledge-Aware Hierarchical Attention Network for Fake News detection on Social Media. Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25 - 29, 2022. https://doi.org/10.1145/3487553.3524664
7. Dun, Y., Tu, K., Chen, C., Hou, C., Yuan, X.: KAN: Knowledge-aware Attention Network for Fake News Detection. Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pp. 81–89. AAAI Press (2021). https://ojs.aaai.org/index.php/AAAI/article/view/16080
8. Shu, K., Cui, L., Wang, S., Lee, D., Liu, H.: dEFEND: Explainable Fake News Detection. In: Teredesai, A., Kumar, V., Li, Y., Rosales, R., Terzi, E., Karypis, G. (eds.) Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019, pp. 395–405. ACM (2019). https://doi.org/10.1145/3292500.3330935
9. Kaliyar, R., K., Goswami, A., Narang, P.: FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimedia Tools and Applications, Volume 80, number 8, pp. 11765–11788. Springer Verlag (2021). https://doi.org/10.1007/s11042-020-10183-2
10. Ahuja, N., Kumar, S.: Fusion of Semantic, Visual and Network Information for Detection of Misinformation on Social Media. Cybernetics and Systems, pp. 1–23. Taylor & Francis (2022)
11. 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, ser. KDD '18, New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 849–857, isbn: 978-1-4503-5552-0. https://doi.org/: 10.1145/3219819.3219903.
12. Cui, L., Wang, S., Lee, D.: SAME: Sentiment-aware multi-modal embedding for detecting fake news. In Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Vancouver British Columbia Canada: ACM, Aug. 2019, pp. 41–48, isbn: 978-1-4503-6868-1. https://doi.org/10.1145/3341161.3342894 .
13. Ayetiran, E. F., Özgöbek, Ö.: An Inter-Modal Attention-Based Deep Learning Framework Using Unified Modality for Multimodal Fake News, Hate Speech and Offensive Language Detection. Available at SSRN: https://ssrn.com/abstract=4504061 or http://dx.doi.org/10.2139/ssrn.4504061 (2023). https://doi.org/10.2139/ssrn.4504061
14. Lommatzsch, A., Kille, B., Özgöbek, Ö., Zhou, Y., Tešić, J., Bartolomeu, C., Semedo, D., Pivovarova, L., Liang, M., Larson, M.: Addressing the depiction gap with an online news dataset for text-image rematching. Proceedings of the 13th ACM Multimedia Systems Conference, pp. 227–233. (2022)
15. Martel, C., Pennycook, G., Rand, D. G.: Reliance on emotion promotes belief in fake news. Cognitive Research: Principles and Implications, 5(1), 1–20. https://doi.org/10.1186/s41235-020-00252-3. (2020)
16. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: In Learning

Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event. Meila, M., Zhang, T. (eds), pp. 8748–8763. PMLR (2021). http://proceedings.mlr.press/v139/radford21a.html

17. Devlin, J., Chang, M., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. IN Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423
18. He, K., Zhang, X., Ren, S. Sun, J.: Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 770–778. IEEE (2016). https://doi.org/10.1109/CVPR.2016.90
19. Pennington, J., Socher R., Manning, C. D.: Glove: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar,A meeting of SIGDAT, a Special Interest Group of the ACL. Moschitti, A., Pang, B., Daelemans, W. (eds), pp. 1532–1543. ACL (2014). https://doi.org/10.3115/v1/d14-1162
20. Yamada, I., Asai, A., Sakuma, J., Shindo, H., Takeda, H., Takefuji, Y., Matsumoto, Y.: Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations", Oct, 2020, Online, pp. 23–30. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.4
21. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track. Bengio, Y., LeCun, Y. (eds).ICLR (2015).
22. Hu, L., Wei, S., Zhao, Z., Wu, B.: Deep learning for fake news detection: A comprehensive survey, AI Open, volume 3, 2022, pp.133–155. https://doi.org/10.1016/j.aiopen.2022.09.001
23. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: FakeNewsNet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media, Big Data, volume 8, issue 3, 2020, pp. 171–188. https://doi.org/10.1089/big.2020.0062