DeBERTNeXT: A Multimodal Fake News Detection framework

Kamonashish Saha[1][0009−0005−3264−9992] and Ziad Kobti[1][0000−0001−9503−9730]

University of Windsor, Ontario N9B 3P4, Canada
{saha91, kobti}@uwindsor.ca

Abstract. With the ease of access and sharing of information on social media platforms, fake news or misinformation has been spreading in different formats, including text, image, audio, and video. Although there have been a lot of approaches to detecting fake news in textual format only, multimodal approaches are less frequent as it is difficult to fully use the information derived from different modalities to achieve high accuracy in a combined format. To tackle these issues, we introduce DeBertNeXT, a multimodal fake news detection model that utilizes textual and visual information from an article for fake news classification. We perform experiments on the immense Fakeddit dataset and two smaller benchmark datasets, Politifact and Gossipcop. Our model outperforms the existing models on the Fakeddit dataset by about 3.80%, Politifact by 2.10% and Gossipcop by 1.00%.

Keywords: Multi-modal · Fake News · DeBERTa · ConvNext.

1 Introduction

Fake news gained attention after the 2016 US presidential election when false news spread mainly through social media, which is the primary news source for 14% of Americans [1]. There are fact-checking websites which include Politifact, AltNews, Fact Check, as well as expert-based and crowdsourced methods to verify the news. However, manual methods are time-consuming and inefficient given the vast amount of news generated globally. Automatic methods for detecting false news are becoming increasingly popular. Detecting fake news using only textual content has been heavily researched, but articles with images are retweeted 11 times compared to the ones with only text [7]. Therefore, combining data from various modalities is crucial for classification. Researchers proposed models for multimodal detection, but fake news is usually classified as a binary problem as it is a distortion bias which itself is illustrated as a binary problem [18]. It is challenging to attain good accuracy on large datasets and utilize all the features from different modalities. Hence, we propose DeBERTNeXT which is trained on Fakeddit, Politifact, and Gossipcop datasets where our model outperforms other state-of-the-art models in terms of accuracy and other metrics.

In this paper, we propose DeBERTNeXT which is a transfer learning-based architecture that utilizes textual and visual features for classifying news as real or

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_36

fake. The representations from both modalities are concatenated for classification and it is not dependent on sub-tasks or domain-specific. We trained and tested our model on Fakeddit, Politifact, and Gossipcop datasets where we achieved better classification results in terms of accuracy, precision, recall, and F1 scores compared to other models to the best of our knowledge.

2 Literature Review

Knowledge-graph-based approaches can verify the veracity of the main statements in a news report [18]. Such methods are inefficient for extensive data and depend on an expert to assign the truthfulness of the news [18]. A significant amount of research has been focused on the textual content which uses BERT and RoBERTa [19]. Unimodal approaches are not suitable for multimodal formats and new architectures were subsequently proposed. One baseline multimodal architecture includes SpotFake [22], which uses BERT for learning text features and pre-trained VGG-19 for the image features. SpotFake+[21] was also introduced later, using pre-trained XLNet and VGG-19 for the combined image and text classification. Another architecture was named Event Adversarial Neural Networks for Multi-Modal Fake News Detection (EANN), which was an end-to-end model for the event discriminator and false news detection [23]. The text representation was attained using CNN, and VGG-19 was used for image representation, where both were later concatenated for classification. Similarly, another model using VGG-19 and bi-directional LSTMs for textual features was introduced, which the authors named Multimodal Variational Autoencoder for Fake News Detection (MVAE) [10]. In addition, FakeNED [15] was introduced, which utilizes finetuned BERT and VGG-19 for binary classification. The baseline multimodal models give good accuracy, but they mainly concentrate on using BERT and VGG architecture, which has some shortcomings of their own. Moreover, the discussed models are trained and tested on small datasets, and some depend on events or require additional preprocessing steps of the dataset. Hence, to tackle these shortcomings, we propose our framework that uses DeBERTa and ConvNeXT which we later fine-tune to train and test on the three different datasets.

3 Methodology

Given a set of n news articles which includes text and image content (multimodal), the data as a collection of a text-image tuple can be represented as:

$$A = (A_i^T, A_i^I)_{i=1}^n \tag{1}$$

Where, A_i^T represents textual content, A_i^I represents the image content and n represents the number of news articles. Since we are considering fake news detection as a binary classification problem, we represent labels as $Y = \{0, 1\}$ where 0 represents fake and 1 represents real or true news. For a given set of

DOI: 10.1007/978-3-031-36021-3_36

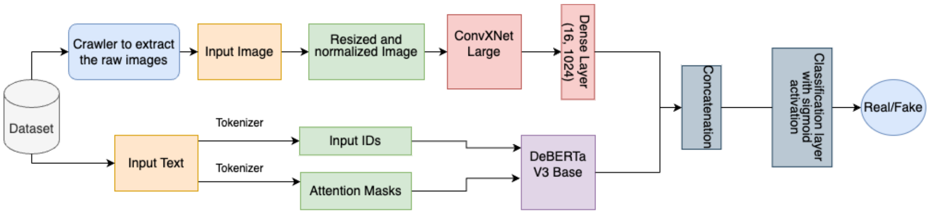

Fig. 1: Model Architecture

news records A, a set of features can be extracted from the textual and image information as represented by X_i^T and X_i^I. The objective of multimodal fake news detection is to create a model $F : \{X_i^T, X_i^I\} \epsilon X \rightarrow Y$ in order to deduce the potential labels in news articles A. Hence, the task of the model is to detect whether the news article A is either fake or real such that:

$$f(A) = \begin{cases} 0, & \text{if } A \text{ is fake news} \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

DeBERTNeXT uses Transformer and Convolutional models for text and images of a particular record. For attaining the textual features, the **DeBERTa** (**D**ecoding-**e**nhanced **BERT** with disentangled **a**ttention) model [5] was used which is based on BERT [3] and RoBERTa [12] models. DeBERTa improves these state-of-the-art models by incorporating disentangled attention mechanism and enhanced mask decoder. For the image part, we have used **ConvNeXT** [13] which is a convolutional model (ConvNet) trained on the large ImageNet dataset.

Image Inputs: The proposed model consists of the DeBERTa V3 base model for the text inputs and ConvNeXT Large cased [13] for the image input. The ConvNeXT model was configured with the proposed model for transfer learning. All the weights of the ConvNeXT model were used except for the last classification layer from the ImageNet pre-trained version. The last layer of the ConvNeXT model with 1536 dimension output was replaced with a fully connected linear dense layer with 1024 output nodes and was found to give the best results. This was chosen based on several experimentations and was set as a hyperparemeter.

Textual Inputs: The improved DeBERTa V3 [4] consists of 12 layers and a hidden size of 768. It consists of 86M backbone parameters with a vocabulary containing 128K tokens which introduce 98M parameters in the Embedding Layer and was trained using 160 GB data. It takes input IDs and attention masks as inputs and outputs a 768-dimension-long tensor. For both images and text, a batch size of 16 was used.

Concatenation and Classification: The output from the DeBERTa V3 model is concatenated with the ConvNeXT large output along the first axis. The concatenation layer takes 768 dimension output from the DeBERTa V3 model and 1024 dimension output from the ConvNeXT model, producing 1792 dimension output. A connected layer with a sigmoid activation function generates the final classification output. The last layer takes 1792 dimension output from the concatenation layer as input and outputs a value ranging from 0 to 1.

4 Experimentation and Results

Our model was trained on Fakeddit, Politifact, and Gossipcop datasets. Various experiments were conducted manually where we used different hyperparameters to get the best set in order to produce the best performance. The image size was 224x224, and the maximum text length varied by dataset. The Fakeddit dataset was set to a maximum text length of 48 words, while Politifact and Gossipcop to 32 words which were set after analyzing the mean and maximum sequence length in the text inputs. We split the training data into an 80:20 train-validation set and kept the test set separate until training was complete.

Only normalization layers and bias were excluded from weight decay while setting up the AdamW optimizer. For the Fakeddit dataset, a batch size of 16 and the maximum learning rate chosen was $3e^{-}6$ and was scheduled with the help of the scheduler. The model was trained for 4 epochs and evaluated on the validation set after every epoch. However, for the Politifact and the Gossipcop dataset, the maximum learning rate was chosen to $2.5e^{-}6$ and was trained for 6 epochs. This was found to be the optimum based on several experiments. Finally, once the training was complete, it was tested on the test set, which was 20% of the respective datasets. The experimentation was carried out in the Google Colaboratory platform using Tesla T4 GPU, and the code is publically available which can be found at https://github.com/Kamonashish/DeBERTNeXT.

Dataset: A crawler was developed to download images from the URLs in the dataset. After filtering GIFs, broken, and distorted images, we extracted 154,644 news records for the Fakeddit dataset [14], 304 for the Politifact dataset, and 8,008 for the Gossipcop dataset. We only used records with both usable text and image data that would refer to the same unique ID of that record.

Data Processing: All the images were extracted using the crawler with the help of the Python libraries: Beautiful Soup and urllib. Once the filtered and refined images were attained, the selected images were reshaped and normalized later during the training phase. Only the URLs were removed for the text since different transformer architectures accept tokens instead of plain English words.

Dataset Pipeline: The data loader pipeline increases loading time significantly as compared to directly loading data to RAM but is necessary as required in case of large datasets when there are RAM constraints. The dataset pipeline accepts the image file path, the text and the label for a particular record in the dataset and processes it to tensors which are accepted by the model. All the images are first loaded from their file path and then reshaped to (224, 224, 3). The reshaping is followed by normalizing the images then converted to tensors. All the text inputs are passed through a model-specific tokenizer which is obtained from the HuggingFace library [25]. All the textual content in the record is tokenized to the maximum length and the text beyond the maximum length is truncated. Special tokens such as [CLS], [SEP] and [PAD] are used. The tokenizer output is in the form of tensors as `input_ids` and `attention_masks`. The `input_ids` are the tokenizer-processed tokens extracted from text, whereas `attention_masks` are tensors of 0 and 1 that serve the purpose of allowing the model to use attention on selected tokens. The `input_ids` and `attention_masks` are given as output for the text inputs of the DeBerta V3 model. At last, the `2_way_label`, which is provided as 0 or 1 in the dataset for fake or real news is also converted to a tensor and given as output to the dataset pipeline.

After training was completed, the model was evaluated on the test set on a variety of classification and evaluation metrics which include accuracy, precision, recall and F1 score. The ROC AUC score, True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) were also noted. Based on our literature review, these were used in calculations for accuracy, recall, precision and F1 score to compare with other models. The best result of three experiments was considered. The model's accuracy and loss plot per epoch for each dataset can be found in .ipynb files in the GitHub repository.

Table 1: Comparison with other models on Fakeddit dataset. (-) indicates that results were not published

Models	**Accuracy**	**Fake News**			**Real News**		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
VGG 19 + Text-CNN [17]	0.804	0.838	0.749	0.791	0.704	0.728	0.716
VQA* [2]	0.631	0.712	0.512	0.596	0.590	0.693	0.637
NeuralTalk [27]	0.612	0.698	0.610	0.651	0.612	0.712	0.658
att-RNN [6]	0.745	0.798	0.637	0.708	0.627	0.713	0.667
EANN [23]	0.699	0.750	0.628	0.684	0.648	0.720	0.682
MVAE [10]	0.784	0.789	0.699	0.741	0.702	0.717	0.709
FakeNED [15]	0.878	-	-	-	-	-	-
CNN for text and image [16]	0.870	-	-	-	-	-	-
DeepNet [8]	0.864	-	-	-	-	-	-
DistilBERT + VGG 16 [9]	0.604	-	-	-	-	-	-
DeBERTNeXT	**0.912**	**0.910**	**0.950**	**0.930**	**0.917**	**0.854**	**0.884**

From Table 1, the authors fused the outputs from Text-CNN with VGG-19 to attain an overall accuracy of 0.804 [17]. Moreover, they modified and trained the VQA [2], NeuralTalk [27], att-RNN [6], EANN [23] and MVAE [10] on the Fakeddit dataset for binary classification. We include their results for comparison. The authors [17] modified VQA by using a binary class layer as the final layer and renamed it as VQA*. In FakeNED [15], after the textual and visual features were extracted, a step was added where a single one-dimensional tensor was passed to fully connected layers for binary classification, where it attained an accuracy of 0.878 and an F1 score of 0.910. Compared with FakeNED, our model outperforms the accuracy by 3.80% and achieves an F1 score of 0.912 by weighted average. CNN was used for both text and images in [16] where both the feature vectors attained after the convolutional layer were passed through two dense layers with ReLU non-linear activation before being concatenated. In the end, the logsoftmax function is applied to attain a micro-average accuracy of 0.870, precision of 0.880, recall of 0.870 and F1 score of 0.870. Similarly, DeepNet [8], which has ReLU as an activation function and softmax function for the final output layer, attains a precision of 0.894, recall of 0.850, an F1 score of 0.872 and an accuracy of 0.864. As compared to the macro-average, our model also outperforms them as we attain an accuracy of 0.912, precision of 0.913, recall of 0.902 and F1 score of 0.917, respectively.

Table 2: Comparison with other models on Politifact and Gossipcop Dataset. (-) indicates that the results were not published

Models	**Dataset: Politifact**				**Dataset: Gossipcop**			
	Accuracy	**Precision**	**Recall**	**F1**	**Accuracy**	**Precision**	**Recall**	**F1**
SpotFake+ [21]	0.846	-	-	-	0.856	-	-	-
SAFE [29]	0.874	0.889	0.903	0.896	0.838	0.857	**0.937**	0.859
Cross-Domain Detection [20]	0.840	0.836	0.831	0.835	0.877	0.840	0.832	0.836
att-RNN [6]	0.769	0.735	**0.942**	0.826	0.743	0.788	0.913	0.846
CMC [24]	0.894	-	-	-	0.893	-	-	-
EANN [23]	0.740	-	-	-	0.860	-	-	-
MVAE [10]	0.673	-	-	-	0.775	-	-	-
SpotFake [22]	0.721	-	-	-	0.807	-	-	-
SceneFND [28]	0.832	-	-	-	0.748	-	-	-
DeBERTNeXT	**0.913**	**0.921**	0.914	**0.915**	**0.902**	**0.914**	0.902	**0.906**

Most of the multimodal models use transfer learning to improve the accuracy of detection. From Table 2, SpotFake+ [21], SpotFake [22], CMC [24] use VGG19 for image content and a transformer-based model such as XLNet [26] and BERT for the textual content. XLNet has a potential for bias [11] and other disadvantages such as more computational cost and longer training time. The authors of SpotFake+ have attained an accuracy of 0.846 for the Politifact dataset and 0.856 on the Gossipcop dataset, which is less as compared to our model. This can be because DeBERTa has some notable advantages over the XLNet model for the textual representations which include higher efficiency, better performance on downstream tasks and improved masking strategy, pre-training techniques and flexibility in model size.

Similar to the textual aspect, our model has notable advantages as it uses ConvXNet over the commonly incorporated VGG 19 used in SpotFake, CMC and Spotfake+ for the image content. Advantages include improved accuracy, scalability, improved regularization, etc. The authors of SpotFake+ have trained the SpotFake, MVAE and EANN models on the Politifact and Gossipcop datasets. We have used those results for a comparison with our model and it can be seen that our model outperforms all of them in terms of accuracy. SAFE [29] uses Text-CNN architecture for both image and text, but transformer models like DeBERTa perform better due to their bidirectional attention mechanism which allows it to better capture contextual information. SAFE's methodology beats our model's recall result on the Gossipcop dataset, but our model performs better on all other metrics.

5 Conclusion

We presented a new multimodal framework that can detect fake news using images and text in a large and small datasets. Our model utilizes the combined power of the transformer model and the ConvNeXT architecture for textual and image content. As per our literature review, our model achieves a better result than other models trained on the same dataset for binary classification. In future, we plan to incorporate more modalities such as audio, video and extract features to combine them with textual content for a complete all-modal fake news detection framework while achieving satisfying classification results.

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. Journal of economic perspectives **31**(2), 211–36 (2017)
2. 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. pp. 2425–2433 (2015)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint:1810.04805 (2018)
4. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
5. He, P., Liu, X., Gao, J., Chen, W.: Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint:2006.03654 (2020)
6. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 795–816 (2017)
7. Jin, Z., Cao, J., Zhang, Y., Zhou, J., Tian, Q.: Novel visual and stat. image features for microblogs news verification. IEEE trans on multimedia **19**(3), 598–608 (2016)
8. Kaliyar, R.K., Kumar, P., Kumar, M., Narkhede, M., Namboodiri, S., Mishra, S.: Deepnet: an efficient neural network for fake news detection using news-user engagements. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). pp. 1–6. IEEE (2020)
9. Kalra, S., Kumar, C.H.S., Sharma, Y., Chauhan, G.S.: Multimodal fake news detection on fakeddit dataset using transformer-based architectures. In: Machine Learning, Image Processing, Network Security and Data Sciences: 4th International Conference, MIND 2022, Virtual Event, January 19–20, 2023, Proceedings, Part II. pp. 281–292. Springer (2023)
10. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: Mvae: Multimodal variational autoencoder for fake news detection. In: WWW. pp. 2915–2921 (2019)
11. Kirk, H.R., Jun, Y., Volpin, F., Iqbal, H., Benussi, E., Dreyer, F., Shtedritski, A., Asano, Y.: Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular gen. language models. Advances in NIPS **34**, 2611–2624 (2021)
12. 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 preprint:1907.11692 (2019)

13. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
14. Nakamura, K., Levy, S., Wang, W.Y.: r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. arXiv preprint:1911.03854 (2019)
15. Sciucca, L.D., Mameli, M., Balloni, E., Rossi, L., Frontoni, E., Zingaretti, P., Paolanti, M.: Fakened: A dl based-system for fake news detection from social media. In: Int. Conf. on Image Analysis and Processing. pp. 303–313. Springer (2022)
16. Segura-Bedmar, I., Alonso-Bartolome, S.: Multimodal fake news detection. Information **13**(6), 284 (2022)
17. Shao, Y., Sun, J., Zhang, T., Jiang, Y., Ma, J., Li, J.: Fake news detection based on multi-modal classifier ensemble. In: Proceedings of the 1st International Workshop on Multimedia AI against Disinformation. pp. 78–86 (2022)
18. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD exp. newsletter **19**(1), 22–36 (2017)
19. Shushkevich, E., Alexandrov, M., Cardiff, J.: Bert-based classifiers for fake news detection on short and long texts with noisy data: A comp. analysis. In: International Conference on Text, Speech, and Dialogue. pp. 263–274. Springer (2022)
20. Silva, A., Luo, L., Karunasekera, S., Leckie, C.: Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 557–565 (2021)
21. Singhal, S., Kabra, A., Sharma, M., Shah, R.R., Chakraborty, T., Kumaraguru, P.: Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In: Proceedings of the AAAI. vol. 34, pp. 13915–13916 (2020)
22. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). pp. 39–47. IEEE (2019)
23. 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. pp. 849–857 (2018)
24. Wei, Z., Pan, H., Qiao, L., Niu, X., Dong, P., Li, D.: Cross-modal knowledge distillation in multi-modal fake news detection. In: ICASSP 2022-2022 IEEE Int. Conf. on Acoustics, Speech and Signal Proc. (ICASSP). pp. 4733–4737. IEEE (2022)
25. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. pp. 38–45 (2020)
26. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems **32** (2019)
27. Yu, E., Sun, J., Li, J., Chang, X., Han, X.H., Hauptmann, A.G.: Adaptive semi-supervised feature selection for cross-modal retrieval. IEEE Transactions on Multimedia **21**(5), 1276–1288 (2018)
28. Zhang, G., Giachanou, A., Rosso, P.: Scenefnd: Multimodal fake news detection by modelling scene context information. Journal of Information Science p. 01655515221087683 (2022)
29. Zhou, X., Wu, J., Zafarani, R.: : Similarity-aware multi-modal fake news detection. In: Advances in KDDM: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II. pp. 354–367. Springer (2020)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_36