FakeRevealer: A Multimodal Framework for Revealing the Falsity of Online Tweets Using Transformer-Based Architectures

Sakshi Kalra[1], Yashvardhan Sharma[1], Priyansh Vyas[1] and Gajendra Singh Chauhan[2]

[1]*Department of CSIS, BITS Pilani, Pilani, 333031, Rajasthan, India*

[2]*Department of HSS, BITS Pilani, Pilani, 333031, Rajasthan, India*

Keywords: Natural Language Processing, Deep Learning, Neural Networks, Transformer-Based Architectures, Multimodal Analysis, Social Media Analytics.

Abstract: As the Internet has evolved, the exposure and widespread adoption of social media concepts have altered the way news is formed and published. With the help of social media, getting news is cheaper, faster, and easier. However, this has also led to an increase in the number of fake news articles, either by manipulating the text or morphing the images. The spread of fake news has become a serious issue all over the world. In one case, at least 20 people were killed just because of false information that was circulated over a social media platform. This makes it clear that social media sites need a system that uses more than one method to spot fake news stories. To solve this problem, we've come up with FakeRevealer, a single-configuration fake news detection system that works on transfer learning based techniques. Our multi-modal archutecture understands the textual features using a language transformer model called DistilRoBERTa and image features are extracted using the Vision Transformer (ViTs) that is pre-trained on ImageNet 21K. After feature extraction, a cosine similarity measure is used to fuse both the features. The evaluation of our proposed framework is done over publicly available twitter dataset and results shows that it outperforms current state-of-art on twitter dataset with an accuracy of 80.00% which is 2.23%more, that than the current state-of-art on twitter dataset.

1 INTRODUCTION

Our modern world is becoming increasingly digital as more and more people rely on the Internet for their news, entertainment, and interpersonal needs. Online social networks (OSNs) like Facebook, Twitter, etc. are at the center of the current wave of digitalization in society. Online social networks (OSNs) provide a means for people to communicate, share ideas, and keep up with current events; as a result, they have become an integral part of many people's daily routines(Lu and Li, 2020),(Grimme et al., 2017). However, it has also resulted in a rapid increase in the number of "fake news" articles, which are news articles that contain intentionally false information. Typically, these news articles are produced through the manipulation of images, text, audio, and video. Since the 2016 US presidential elections, fake news and the spread of misinformation have dominated the news cycle. Some news stories say that Russia has made a lot of fake accounts and social media bots to spread false information during the elections (Lewandowsky et al., 2017). False information is spread widely at the expense of both society and the individual. At first, this kind of fake news could change or even destroy the balance of truth in the news ecosystem. People are forced to accept wrong or skewed ideas that they would normally reject because of the way fake news works (Asghar et al., 2020). The effects of fake news persist in how people interact with and respond to legitimate news. False news can hurt people, so it's important to make a system that can automatically spot it when it shows up on social media. But there are some hard research questions about how to spot fake news on different social platforms. Identification of the source of origin or uploading of the specific news or data on the social network, understanding the actual intention or meaning of the data uploaded, assessing the data's level of authenticity and validity, and coming to a conclusion about whether it is real or fake are just a few of the research problems that have been noted in this regard. Identifying false news is a difficult task because it involves overcoming a number of challenges. The most challenging aspect of detecting fake news is verifying the reliability of the information being examined. Simply put, a "fact"

Kalra, S., Sharma, Y., Vyas, P. and Chauhan, G.
FakeRevealer: A Multimodal Framework for Revealing the Falsity of Online Tweets Using Transformer-Based Architectures.
DOI: 10.5220/0011889800003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 956-963
ISBN: 978-989-758-626-2; ISSN: 2184-4313

is a basic idea constructed from anything that has ever happened in the past, somewhere, and ultimately with or to someone. It does not seem likely that computers will be able to understand the significance of information if they are allowed to decide on their own who receives what information, when it is delivered, and how. This matters because a lot of content on social media relies on the same method of description. As a result, journalistic criteria must be gathered.

As an alternative definition, determining whether or not a news article is fake involves determining how reliable it is. Fact-checking is one way to stop the spread of fake news. Expert-based fact-checking is very accurate but can't be used on a large scale. Crowdsourced fact-checking, on the other hand, is less likely to be accurate but can be used on a large scale. Thus, the era of human-powered fake news detection is over, making way for automated systems (Zhou et al., 2019). There are numerous fact-checking websites available for checking the veracity of online content. These include sites like PolitiFact, BuzzFeed, Snopes, and GossipCop. The World Health Organization (WHO) designated the virus in early January 2020 ”Coronavirus 2 (SARS-CoV-2) and the syndrome coronavirus disease (COVID-19)”. The WHO has made all data and warnings about COVID-19 and the virus public ”Information Epidemic”. The term ”infodemic” refers to a sickness that spreads false information. It's difficult to verify the reliability and veracity of internet shared data, especially when it comes to a terrible disease that threatens humanity. Buzzfeed.com is a digital media, news, and entertainment company based in the United States and helps in the fact-checked assertions about the Coronavirus as shown in Figures 1 and 2.

Fake identification using manual features or single-modal deep learning features has been the subject of prior research. The problem is that it doesn't take into account the fact that tweets often contain more than one type of media. Tweets with images and videos, like GIFs and videos, may get more attention from users than text-only tweets. In order to solve the problem described above, we came up with the idea of a multi-modal fusion architecture called FakeRevealer. This design combines the text and visual content found in tweets in order to deliver a combined model of FND.

The proposed research seeks to create reliable models for a fake news detection system that can help journalists and regular people spot and dismiss false stories.

1. One goal is to look into the prevalence of deceptive visuals in social media and other multimodal systems that mix text and images.

2. Second, we aim to create a model that is both effective at spotting fake news and capable of capturing the shallow dependency relationships between visual and textual content using techniques from the field of transformer-based approaches.

This paper is structured as follows: Section 2 provides an overview of relevant prior work, while Sections 3-4 present the multi-modal datasets used in this investigation and the proposed model architecture and its specifics, respectively. In Section 5, the experimental details of this work are explained, and in Section 6, the work as a whole is summed up.

Figure 1: Fact-Checked Claims associated to COVID-19 by Buzzfeed.com[1].

Figure 2: Fact-Checked audio and video related to COVID-19 by Buzzfeed.com[2].

2 RELATED WORK

Fake News Detection is a binary classification problem that attempts to determine whether information is genuine or manipulated. Most traditional work is all about analyzing text to do things like figure out how someone feels or find fake news. (Conroy et al., 2015) uses a hybrid method that combines machine learning, linguistic clues, and network-based behavioral data. (Pérez-Rosas et al., 2017) uses SVM with five linguistic feature cross validations and focuses on linguistic feature-based approaches. (Pan et al., 2018) employed knowledge graphs to enhance the truth analysis. These graphs are used to extract information about entity relationships from the data. Since neural networks came along in the second decade of this century, deep learning techniques have been used in a lot of different ways. The temporal relationship between words in a sentence is determined by the recurrence neural network-based system in (Ma and Hovy, 2016). However, one of its shortcomings is that it struggles with long phrases. Chen et al. used a self-attention-based configuration to solve the problem (Huang et al., 2022). The researchers also discovered that visual content receives more attention from news readers than textual content, resulting in a stronger impact of the content on people (You et al., 2016). GANs were used by (Marra et al., 2018) to detect fake images, and the splicing technique was used to identify these kinds of images. The goal of (Steinebach et al., 2019) is to automatically recognize photomontages using feature detection. The methods covered above are unimodal and concentrate on either text-based features or visual features. But with social media, it is necessary to pay attention to both modalities. Researchers extracted both the feature list and combined them to create a single unit for multimodal approaches.

(Wang et al., 2018) created an end-to-end model for detecting fake news called Event Adversarial Neural Networks for Multi-Modal Fake News Detection (EANN). They have two parts to their model: text and images. Text representation was created using the CNN model, whereas image representation was taken from the VGG-19. Their model has an accuracy of 64.8% on the Twitter dataset and 79.5% on the Weibo dataset. Multimodal Variational Autoencoder for Fake News Detection (MVAE), a similar type of architecture, was also developed by (Khattar et al., 2019). Text representation was extracted using a bi-directional LSTMs network, while image representation was once more extracted from VGG-19. The modal achieves an accuracy of 74.5% on the Twitter dataset and 82.4% on the Weibo dataset. (Singhal et al., 2019) proposed the SpotFake system and concentrated on multimodal fake news detection. The textual and visual components of an article serve as the foundation for SpotFake. Singhal et al. used the state-of-the-art BERT for textual representation to include contextual information, while for image features they used the VGG-19 pre-trained on ImageNet dataset. The modal performs with an accuracy of 77.77% on twitter dataset and 89.23% on weibo dataset. Several authors have also come up with models for spotting fake news, which they have tested using the Fakeddit dataset. (Kirchknopf et al., 2021) uses the Fakeddit dataset to perform fake news detection using four different modalities, namely the news content, comments, images, and metadata. To identify fake news, (Shao et al., 2022) proposed an ensemble method. To do this, they first built two unimodals, one on text and the other on an image, and then built a multi-modal after using all three as inputs to the ensemble classifier.

All of the models mentioned above did well in the multimodal fake news detection FND, but there is room for improvement in the measure of similarity between text and visual features for the Twitter dataset. And for the same, we suggest FakeRevealer, a cutting-edge standalone multimodal fake news detection tool.

3 DATASET USED

The dataset repository consists of one dataset that is from the Twitter media domain (Boididou et al., 2015) and was released for a challenge at Verifying Multimedia Use at MediaEval on multimediaeval.org. The challenge was to figure out if the information in the post was a good representation of reality or not. In this dataset, each entry consists of an article that has a text and an image associated with it. The training sample has 11,663 unique samples and 342 unique images, while the test set is made up of 3,755 Twitter news tweets. For this study, we only looked at real and fake labels and left out records that were humorous. Table 1 lists the dataset statistics used for the proposed work.

Table 1: Dataset Statistics used for the Proposed Work.

Dataset	Real	Fake	Modality	Source
TwitterMediaEval2015	4921	6742	Text + Image	Github

4 PROPOSED METHODOLOGY

The collection of data is the first step in the model's creation. The tweet's text and image content make up the multi-modal model's input. The Fake News Detection label, which can be either R or F depending on the input to the model, is the final result. The proposed model is made up of three components: a textual component, an image component, and a module for combining different types of information (multimodal component).

4.1 Hyperparameters Statistics

To train machine and deep learning-based algorithms efficiently, hyperparameters are crucial because they directly affect how the training algorithm operates. Therefore, the performance of the model is highly sensitive to these parameters. The number of hyperparameters employed by the suggested multimodal architecture is shown in Table 2.

Table 2: Hyperparameters Emplyoed by the Multimodal Architecture.

Hyperparameter	Value
Dense Layers:	3
Dropout Layers:	1
Dropout rate:	0.2
Loss function:	categorical_loss_entropy
Optimizer:	Adam
Activation function:	softmax
Learning rate:	6e-6
Beta 1:	0.9
Beta 2:	0.99
Epochs:	30
Batch size:	12

4.2 Textual Component

This sub-module is responsible for extracting the contextual text features from the posts. We used a distilled version of pretrained RoBERTa-base which is version of BERT model. BERT stands for Bi-directional Encoder Representation from Transformer (Devlin et al., 2018). It uses a transformer to assign weights to every input and output connection. Previously, models were built to read the text sequentially, i.e., either left-to-right or right-to-left. A robustly optimized BERT approach, RoBERTa, is a retraining of BERT with improved training methodology. RoBERTa takes the Next Sentence Prediction (NSP) task out of BERT's pre-training and adds dynamic masking so that the masked token changes during the training epochs. This makes the training process better. But RoBERTa is too large, so we opted for Distil RoBERTa which is a distilled version of the RoBERTa model. In our proposed model, the training inputs are first encoded using the DistilRoBERTa tokenizer, and then the model is finetuned using the encoding. The output of DistilRoBERTa model is passed through a few dense layers, the last layer is a softmax layer with 2 neurons (Fake and Real) and finally we compiled our neural network model using adam optimizer with a learning rate of 1e-03 as shown in Figure 3.

We have also used a GPT-2 transformer, which is self-supervised and has been trained on a large corpus of English data. The model was mainly trained to predict next word, in GPT-2 inputs are the sequence of words and output are the same sequence of words shifted one token right. The model also uses mask-mechanism internally. But since the model is very large, we haven't yet fully discovered its full potential in our proposed method.

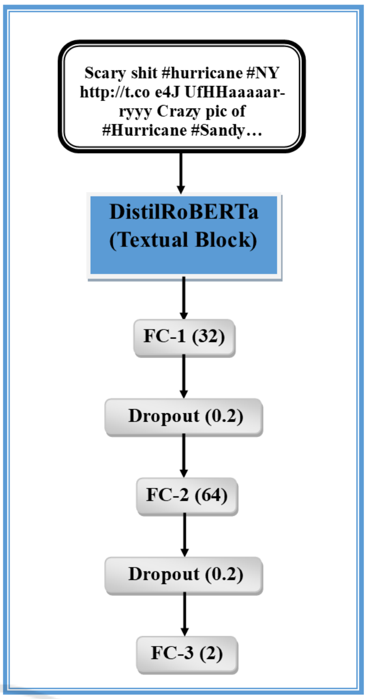
Scary shit #hurricane #NY http://t.co e4J UfHHaaaaar-ryyy Crazy pic of #Hurricane #Sandy…
DistilRoBERTa (Textual Block)
FC-1 (32)
Dropout (0.2)
FC-2 (64)
Dropout (0.2)
FC-3 (2)

Figure 3: Textual Feature Extraction Architecture (Fine-tuned Distill RoBERTa Model).

4.3 Image Component

We used the VGG-19 and the most up-to-date Vision Transformers (ViTs) to pull out features from images. VGG-19 is a convolutional neural network that is trained on images from the ImageNet database and is 19 layers deep. It is made up of 16 layers of convolution: 3 layers that are fully connected, 5 layers of MaxPool, and 1 layer of SoftMax. The Vision transformer, on the other hand, employs a transformer-like structure for image patches as shown in Figure 4. In Vision Transformers (ViTs) an image is split into fixed-size patches; each of them is then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed to a standard Transformer encoder. The standard way to do classification is to add an extra "classification token" that can be learned to the sequence. In the name of each checkpoint, you can see both the patch resolution and the image resolution that were used during pre-training or fine-tuning. The transformer is pre-trained on images from the ImageNet-21K database with a resolution of 224 x 224.

Before applying VGG-19 and ViTs the images have been rescaled to 224 X 224 size and images that cannot be rescaled are being discarded. The trainable layers of VGG-19 are all set to FALSE, and for ViTs, all the layers except last 7 are set to False.

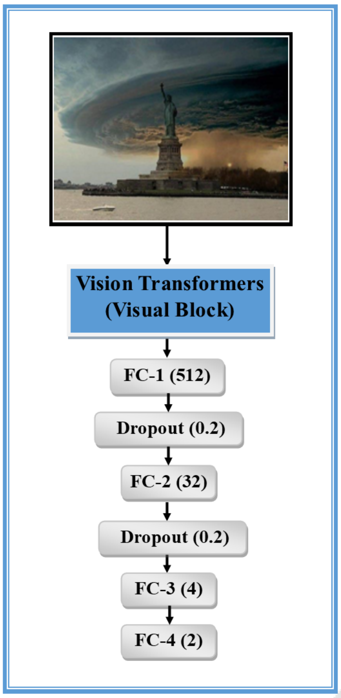

Figure 4: Image Feature Extraction Architecture (Fine-tuned Vision Transformers (ViTs) Model).

4.4 Multimodal Component

For multimodal component, we first preprocess the dataset and then use the techniques we talked about in the textual and image components to extract text and image features at the same time. We then combine the two feature vectors as shown in Figure 5. This fused output layer passed them into a dense-32, fully connected FC-1 layer. The multimodal is compiled using an Adam optimizer with a learning rate of 6e-6 and sparse categorical cross-entropy loss for training the model, followed by a dropout layer with a drop rate of 0.2 and finally the output layer, which uses a softmax activation function.

For pre-processing, first the text records that correspond to the same image are aggregated as shown in Figure 6, then if the size of the record exceeds 500 characters (maximum length BERT can take), the record is split into three halves ranging from [0:200], [200:400], and [400:]. Now the image that is repeated after the transformation is rotated by a 90-degree angle for the split [200:400] and a 180-degree angle for the split [400:..]. and for the split [0:200], the image is kept unchanged. The np.zeros((224,224)) function is used to replace the images that can't be resized with a blank image of size 224 x 224. This keeps the text information that goes with the image. We are processing 200 words at a time using the pre-trained model.

The output of ViTs is 1024 parameters, which are passed to a 768-neuron dense layer, which reduces its size to 768, and then the output of this layer is fused with the output of the Distil RoBERTa layer.

5 EXPERIMENTAL ANALYSIS

Different transformer based architectures are used for the unimodal (Text based analysis, Image based analysis) and multimodal (Text + Image based analysis) in this research work. When making multimodal systems, the main challenge is to keep the features that make each mode unique while combining useful features from many modes.

5.1 Comparative Analysis of Various Transformer-Based Unimodal and Multimodal Architectures

5.1.1 Unimodal (Text Based Results)

Textual-based models are built by removing URLs, punctuation, and stopwords from text data. This data is then fed to a DistilRoBERTa that has already been trained to pull out features. The obtained features are passed to a fully connected dense layer, then a dropout layer removes 20% of neurons, and the final logits are passed to an output layer having a binary class softmax activation function. This model is then trained using the adam optimizer with a learning rate of 6e-6 and a loss function of sparse categorical cross-entropy. In this proposed work, we have tested 3 textual architectures: DistilBERT, DistilRoBERTa and DistilGPT-2, and found that DistilRoBERTa performs better than others. Table 3 shows the accuracy comparison of all three architectures.

Table 3: Unimodal (Textual based Results).

Modality	Model	Accuracy
Unimodal (Text)	DistilBERT	86.28%
Unimodal (Text)	**Distil RoBERTa**	**89.52%**
Unimodal (Text)	DistilGPT-2	67.80%

5.1.2 Unimodal (Image Based Results)

The construction of image-based models starts by pre-processing the image by resizing it to 224x224. In this proposed work, we used three image-based architectures for feature extraction: VGG16 (Qassim et al., 2018), VGG19 (Mateen et al., 2018), and ViTs (Vision Transformer) (Dosovitskiy et al., 2020). Table 4 shows that ViTs outperforms the other two architectures in terms of accuracy.

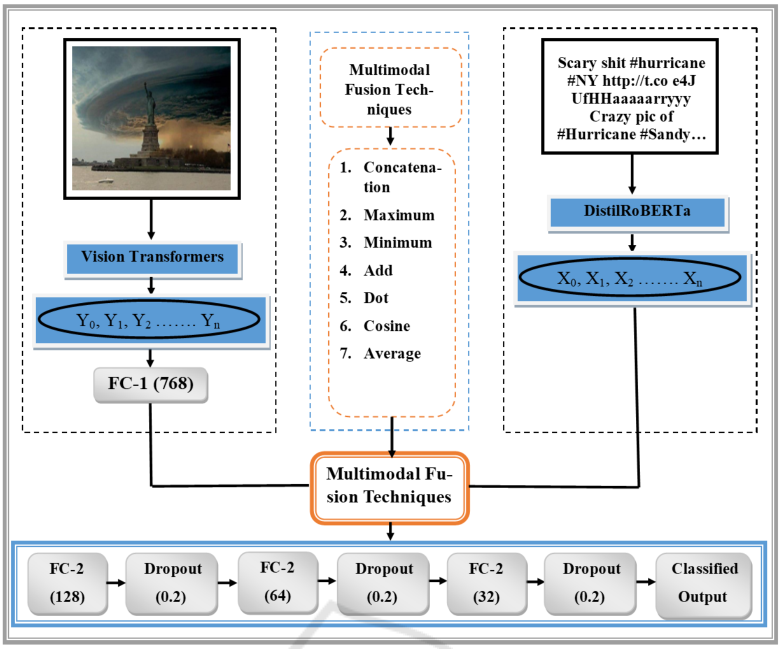

Figure 5: Multimodal Feature Extraction Architecture.

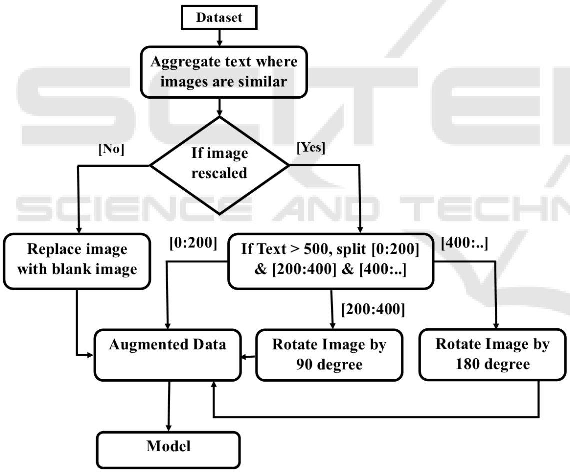
Dataset
Aggregate text where images are similar
If image rescaled
[No]
[Yes]
Replace image with blank image
If Text > 500, split [0:200] & [200:400] & [400:..]
[0:200]
[400:..]
[200:400]
Augmented Data
Rotate Image by 90 degree
Rotate Image by 180 degree
Model

Figure 6: Data Augmentation.

5.1.3 Multimodal (Text+Image Based Results)

The proposed multimodal architecture is a combination of DistilRoBERTa and ViTs, which outperform the current state-of-the-art SpotFake(Singhal et al., 2019) in terms of accuracy. In Table 5, we show how our models compare to the current best models in terms of accuracy, precision, recall, and f1-score. The EANN(Wang et al., 2018) and MVAE(Khattar et al., 2019) both use two model configurations. EANN has two components. The text part used a CNN to generate a text representation from the word embedding vector. and from the VGG-19 model that had already been trained on ImageNet, the image representation was taken. The MVAE model's primary task was to build an auto encoder-decoder model. They used bidirectional LSTMs to get the text representation out of VGG-19, and they also got the image representation out. SpotFake is also a stand-alone configuration model. (Singhal et al., 2019) used the VGG-19 trained on the ImageNet dataset for the image features and the BERT to add context to the textual representation. The VQA (Antol et al., 2015), Neural Talk (Vinyals et al., 2015), and att-RNN (Jin et al., 2017) have also performed well on multimodal analysis. Even though it is a standalone configuration model, the proposed FakeRevealer model does better on the Twitter medieval dataset than EANN, MVAE, and SpotFake.

Table 4: Unimodal (Image based Results).

Modality	Model	Accuracy
Unimodal (Image)	VGG-16	60.83%
Unimodal (Image)	VGG-19	62.50%
Unimodal (Image)	**ViTs**	**66.15%**

The extracted features from both modalities are fused using a variety of fusion techniques, including multiplying, concatenating, and taking the maximum of both features. The cosine function performs more favorably on the multimodal architecture that is being proposed. Concatenation, which is simply concatenating both feature lists; add, which is adding the values of the features; maximum, i.e., selecting the maximum out of both; minimum, which is selecting

Table 5: FakeRevealer vs. Other Multimodal Architectures on the Twitter MediaEval Dataset.

Model	Accuracy	Real(P)	Real (R)	Real (F1)	Fake (P)	Fake (R)	Fake (F1)
EANN	64.8%	81.0%	49.8%	61.7%	58.4%	75.9%	66%
VQA	63.1%	76.5%	50.9%	61.1%	55%	79.4%	65%
Neural Talk	61%	72.8%	50.4%	59.5%	53.4%	75.2%	62.5%
att-RNN	66.4%	74.9%	61.5%	67.6%	58.9%	72.8%	65.1%
MVAE	74.5%	80.1%	71.9%	75.8%	68.9%	77.7%	73%
SpotFake	77.7%	75.1%	90%	82%	83.2%	60.6%	70.1%
FakeRevealer	**80%**	76%	97%	85%	89%	42%	57%

the minimum out of both; average, which is taking the average of both feature lists; dot, which is getting the by-product of both feature lists; and cosine fusion technique, which is selecting the maximum out of both. In Table 6, a comparative analysis of the outcomes of the use of various fusion methods is provided. Figure 7 provides the graphical comparison of all the existing state-of-the-art multimodal architectures with the Proposed model (FakeRevealer).

Table 6: Accuracy Comparison of FakeRevealer Fusion Techniques.

Modality	Fusion Technique	Accuracy
Text + Image	Concatenation	67.20%
Text + Image	Maximum	74.55%
Text + Image	Minimum	67.27%
Text + Image	Add	56.35%
Text + Image	Dot	65.45%
Text + Image	**Cosine**	**80.00 %**
Text + Image	Average	74.55%

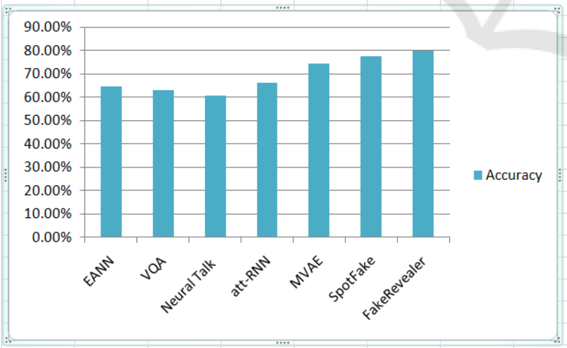

Figure 7: Comparative Analysis of FakeRevealer with various Pre-Existing Multimodal Architectures.

5.2 Error Analysis

We came to the conclusion that the image data is significantly less than the textual data due to the fact that a large number of tweets have been retweeted using the same image, and even after image augmentation, the accuracy of image models is significantly less than that of textual ones. When we combine the features of the two, the total accuracy of the multimodal analysis suffers as a direct consequence of this primary factor. In addition, as we were training the multimodal over cosine similarity algorithm, we saw that the total loss was getting better, but the validation score didn't change.

6 CONCLUSIONS

The proposed text model works well over Twitter-mediaEval dataset with an accuracy of 89.74% and the multi-models works with an accuracy of 80% and there is still room for improvement in the image and multimodal architectures. In future, as we observed while training the multimodal over cosine similarity the overall loss is decreasing but the validation score remains constant. This issue can be further explored. Along with this simple fusion can be accommodated with ensemble classifier and CLIP diffusion to enhance the overall performance of the proposed architecture.

REFERENCES

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Asghar, M. Z., Ullah, A., Ahmad, S., and Khan, A. (2020). Opinion spam detection framework using hybrid classification scheme. *Soft computing*, 24(5):3475–3498.

Boididou, C., Andreadou, K., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Kompatsiaris, Y., et al. (2015). Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3):7.

Conroy, N. K., Rubin, V. L., and Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Grimme, C., Preuss, M., Adam, L., and Trautmann, H. (2017). Social bots: Human-like by means of human control? *Big data*, 5(4):279–293.

Huang, Z., Lv, Z., Han, X., Li, B., Lu, M., and Li, D. (2022). Social bot-aware graph neural network for early rumor detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6680–6690.

Jin, Z., Cao, J., Guo, H., Zhang, Y., and Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Khattar, D., Goud, J. S., Gupta, M., and Varma, V. (2019). Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

Kirchknopf, A., Slijepcevic, D., and Zeppelzauer, M. (2021). Multimodal detection of information disorder from social media. *arXiv preprint arXiv:2105.15165*.

Lewandowsky, S., Ecker, U. K., and Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of applied research in memory and cognition*, 6(4):353–369.

Lu, Y.-J. and Li, C.-T. (2020). Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. *arXiv preprint arXiv:2004.11648*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. (2018). Detection of gan-generated fake images over social networks. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, pages 384–389. IEEE.

Mateen, M., Wen, J., Song, S., and Huang, Z. (2018). Fundus image classification using vgg-19 architecture with pca and svd. *Symmetry*, 11(1):1.

Pan, J. Z., Pavlova, S., Li, C., Li, N., Li, Y., and Liu, J. (2018). Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*.

Qassim, H., Verma, A., and Feinzimer, D. (2018). Compressed residual-vgg16 cnn model for big data places image recognition. In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*, pages 169–175. IEEE.

Shao, Y., Sun, J., Zhang, T., Jiang, Y., Ma, J., and Li, J. (2022). Fake news detection based on multi-modal classifier ensemble. In *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, pages 78–86.

Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., and Satoh, S. (2019). Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Steinebach, M., Gotkowski, K., and Liu, H. (2019). Fake news detection by image montage recognition. In *Proceedings of the 14th International Conference on Availability, Reliability and Security*, pages 1–9.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., and Gao, J. (2018). 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*, pages 849–857.

You, Q., Cao, L., Jin, H., and Luo, J. (2016). Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1008–1017.

Zhou, X., Zafarani, R., Shu, K., and Liu, H. (2019). Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 836–837.