Multimodal Multi-image Fake News Detection

Anastasia Giachanou
Universitat Politècnica de València
Valencia, Spain
angia9@upv.es

Guobiao Zhang
Wuhan University,
Wuhan, China
Universitat Politècnica de València,
Valencia, Spain
zgb0537@whu.edu.cn

Paolo Rosso
Universitat Politècnica de València
Valencia, Spain
prosso@dsic.upv.es

***Abstract*—Recent years have seen a large increase in the amount of false information that is posted online. Fake news are created and propagated in order to deceive users and manipulate opinions and subsequently have a negative impact on the society. The automatic detection of fake news is very challenging since some of those news are created in sophisticated ways containing text or images that have been deliberately modified. Combining information from different modalities can be very useful for determining which of the online articles are fake. In this paper, we propose a multimodal multi-image system that combines information from different modalities in order to detect fake news posted online. In particular, our system combines textual, visual and semantic information. For the textual representation we use the Bidirectional Encoder Representations from Transformers (BERT) to better capture the underlying semantic and contextual meaning of the text. For the visual representation we extract image tags from multiple images that the articles contain using the VGG-16 model. The semantic representation refers to the text-image similarity calculated using the cosine similarity between the title and image tags embeddings. Our experimental results on a real world dataset show that combining features from the different modalities is effective for fake news detection. In particular, our multimodal multi-image system significantly outperforms the BERT baseline by 4.19% and SpotFake by 5.39%.**

***Index Terms*—multimodal fake news detection, multi-image system**

I. Introduction

Recently there has been a huge increase in the amount of inaccurate and manipulated information that is posted online. Fake news usually aim to deceive users and influence opinions by manipulating the textual and multimedia content. Fake news that are propagated online have a negative impact on different aspects of society. For example, some studies support that the outcomes of several elections and referendums such as U.S. presidential elections [3] and the Brexit [2] might have been influenced by fake news shared online in different social media platforms. More recently, a great amount of fake news and misinformation was propagated about the Coronavirus disease (COVID-19). Fake news about the effectiveness of the chloroquine led to an increase of cases of chloroquine drug overdose [4], whereas rumors about the lock down led to panic buying of groceries and paper products in several countries that subsequently had negative consequences on the supply management [22].

Fake news is not a new phenomenon but exists for a long time. However, the access to social media where anyone can post anything in a very easy way has escalated fake news propagation. The detection of fake news is very challenging and even humans are not able to distinguish real from fake content since they usually contain mixture of fake and real information. There are different platforms developed with the aim to raise awareness to the users about misinformation posted online. The annotation of the articles on those platforms is done manually by journalists and other experts who analyse the content of the articles and determine whether it is fake or not. For example, Politifact[1] contains labels for claims that are mainly focused on political news, whereas GossipCop[2] contains annotated articles about celebrities and entertainment.

The huge amount of online misinformation makes the development of a system that can automatically detect fake news a necessity. Early works focused on using textual information extracted from the text of the article, such as statistical text features [5] and emotional information [1], [9], [12]. Although the textual content can be a very important indicator for fake news detection, it is not sufficient when it is used alone. Online articles and posts usually contain more information such as images and social context that can be also useful for fake news detection. To this end, some researchers have proposed systems that use the credibility of the pages that post the news [17] or profile characteristics of the users that shared the post to detect the articles that contain manipulated content [11], [20].

Online news contain also images that usually attract the attention of the users. It is possible that images in fake and real news follow different patterns or that have been modified in order to attract users' attention and make them share them. Hence, it is important that a system also exploits information extracted from the images for effective fake news detection. Visual information can complement the textual one for fake news detection. Some researchers have proposed multimodal systems that combine textual and visual information for determining whether an article or a post is fake or not [21], [26]. These systems are usually based on visual features that are extracted only from one image ignoring that there are posts

[1] https://www.politifact.com/

[2] https://www.gossipcop.com/

with more than one image. However, articles contain more than one image and additional information that could be useful for the detection could be extracted. Different from previous works on multimodal fake news detection, we propose to extract and combine visual features that are extracted from more than one image in cases of articles that have multiple images. The visual features are combined and passed to a Long Short Term Memory (LSTM) layer. Figure 1 shows an example of an article with the title *Does Brad Pitt Have a Rare Disease?*. This article contains three different images.

Finally, our system also uses the similarity between the image and the title of the article which is a type of information that has not been extensively explored in the context of fake news. The text-image similarity can be a valuable indicator especially in the cases that the images are chosen randomly and do not correspond to the article.

The rest of the paper is organised as follows: Section II presents the related work on fake news detection. The multimodal multi-image system is presented in Section III. Section IV provides the experimental setup and details about the collection and the settings. The results of the experiments are presented in Section V whereas Section VI concludes the study.

II. Related Work

Fake news detection has recently received a lot of research attention. Early attempts on fake news detection were mainly focused on information that was extracted from text to capture the different linguistic patterns used in fake and real news. One of the early works was presented by Castillo et al. [5] who explored the effectiveness of various statistical text features, such as count of word and punctuation on information credibility. More recently, Rashkin et al. [18] incorporated various linguistic features extracted with the Linguistic Inquiry and Word Count (LIWC) dictionary [28] such as personal pronouns and swear words into an LSTM network in order to differentiate between credibile and not credible claims, whereas Wang [25] proposed a hybrid convolutional neural network to combine user metadata with text for fake news detection.

Based on the intuition that fake news triggers different emotions compared to real news to the users, some researchers proposed extracting the emotions expressed in the text and they explored their effectiveness on the task of fake news detection. Vosoughi et al. [24] investigated true and false rumours on Twitter and found that false rumours triggered fear, disgust and surprise in their replies, whereas the true rumours triggered joy, sadness, trust and anticipation. Giachanou et al. [12] proposed an LSTM-based neural network that leveraged emotions expressed in the text. They explored three different ways to extract the emotions, two of them were lexicon-based and one was based on a neural network. In their study, Giachanou et al. showed the effectiveness of the emotions expressed in the text on credibility detection. Another work that explored the impact of emotions on fake news detection was presented by Ghanem et al. [9]. Ghanem et al. who proposed to extract the emotions expressed in the text and incorporated them into an LSTM network showed that emotions are useful for the classification of the different types of fake news.

Users can also play an important role in the propagation of fake news since they are the ones that decide to share the fake information intentionally or unintentionally. To this end, some researchers explored the role of users in the detection and propagation of fake news. Shu and Wang [20] performed an analysis of user profiles that share fake or real news. The analysis showed that there are features (e.g., registration time) that are different between users that share fake news and those that share real news. In addition, they examined the effectiveness of those features on fake news detection and showed that combining user profile features with the psycholinguistic characteristics of the document can be very effective for fake news detection. Vo and Lee [23] analysed linguistic characteristics of fact-checking tweets (i.e., tweets that confirm that an article is fake) and also proposed a deep learning framework to generate responses with fact-checking intention. Their analysis showed that the fact-checkers tend to refute fake news and use formal language.

Multimodal fake news detection has also received research attention since the majority of the articles contain one or more images. Fake news usually contain images that are manipulated in sophisticated ways to deceive the users, attract their attention and convince them to share them. Visual information extracted from the images, such as image tags and colour histogram can complement the textual one. Different studies have focused on that and showed that visual features can be an important indicator for fake news detection. Jin et al. [13] proposed several visual and statistical features to characterize different patterns used in fake and real news in order to detect fake news. However, their work was based on hand-crafted features that cannot capture complex distributions of visual content.

More recently, the multimodal approaches that were proposed exploited the advances of deep learning area. Wang et al. [26] proposed the Event Adversarial Neural Networks (EANN) model that consists of two components: the textual and the visual. The textual component was represented by word embeddings generated using Convolutional Neural Network (CNN), whereas the visual one was represented by features that were extracted using the VGG-19 model pretrained on ImageNet [6]. The two representations were then concatenated and fed to two fully connected neural networks, one network was used for event discriminator and the second for fake news classification.

Khattar et al. [14] proposed an end-to-end network, Multimodal Variational Autoencoder (MVAE) model based on bi-directional LSTMs and VGG-19 for the text and image representation respectively. The model consists of three main components, an encoder, a decoder and a fake news detector module. The variational autoencoder is capable of learning probabilistic latent variable models whereas the fake news detector utilizes the multimodal representations obtained from the variational autoencoder to classify posts as fake or not.

Fig. 1: Images of an article with the title *Does Brad Pitt Have a Rare Disease?*

Singhal et al. [21] focused also on multimodal fake news detection and proposed the SpotFake system. SpotFake is based on the textual and visual features of an article. For the textual representation, Singhal et al. used BERT to incorporate contextual information, whereas for the image features, they used the VGG-19 pre-trained on ImageNet dataset. The representations from both the modalities are then concatenated together to produce the desired news vector. Their results showed the importance of combining contextual information and visual features for fake news detection.

Regarding the image-text similarity, there are few works that have explored its effectiveness on fake news detection. Zlatkova et al. [31] explored the effectiveness of text-image similarity in addition to other visual information. However, Zlatkova et al. focused on claim factuality prediction with respect to an image that is a different problem to the one of fake news detection. Zhou et al. [30] proposed the Similarity-Aware FakE news detection method (SAFE) that consisted of three components, the multimodal one, the within modal and the cross-modal similarity extraction. Zhou et al. used neural networks to automatically obtain the latent representation of the textual and visual information based on which a similarity measure was defined between them.

Unlike previous works, we propose visual features that are extracted from multiple images and which we pass then into an LSTM layer to model the sequence information. In addition, we explore the effectiveness of the similarity between text and image that can better capture different patterns used in fake and real news. For the textual component we use BERT that can learn the context of a word based on all of its surroundings, whereas the similarity is calculated using the embeddings of the post's text and the image tags.

III. Methodology

In this section we describe the multimodal multi-image fake news detection system. Our system consists of three different components: the textual, the visual and the semantic component. Figure 2 shows the architecture of our system.

- *Textual:* The content of the post is the most important information for the detection of fake news and has been shown to be useful for a wide range of text classification tasks, from reputation analysis to irony detection and from sentiment analysis to credibility detection [8], [10], [12]. Previous fake news detection approaches have used representations such as bag-of-words and word embeddings [5], [9], [18]. These representations cannot capture the contextual information. On the contrary, our system uses a more sophisticated representation that is based on Bidirectional Encoder Representations from Transformers (BERT) [7]. BERT is applying the bidirectional training of a Transformer, an attention mechanism that learns contextual relations among words in a text. The Transformer includes two separate mechanisms, an encoder and a decoder. The encoder aims to read the text input and the decoder outputs a prediction for the task. Different to previous systems that looked at a text sequence either from left to right or combined left-to-right and right-to-left training, the Transformer encoder reads the entire sequence of words at once. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

In particular, we use the pretrained BERT-Base available on TensorFlow Hub[3]. The padded and tokenized text is passed into the BERT model to receive word vectors of dimension 768. BERT pre-trained model has shown to be effective for improving many natural language processing tasks.

- *Visual:* The visual information can be very useful in case there are different patterns used in fake and real news or there are images that have been manipulated. The novelty of our system is that it combines visual information from more than one image. For the image representation, we use the VGG-16 model that is pre-trained on the visual dataset ImageNet and which contains over 14 million hand-annotated images [6]. Similar to many existing works on visual classification that have also used the VGG16 features [32], we adopt the output of the last layer of the VGG16 to represent the visual features.

[3] https://www.tensorflow.org/hub

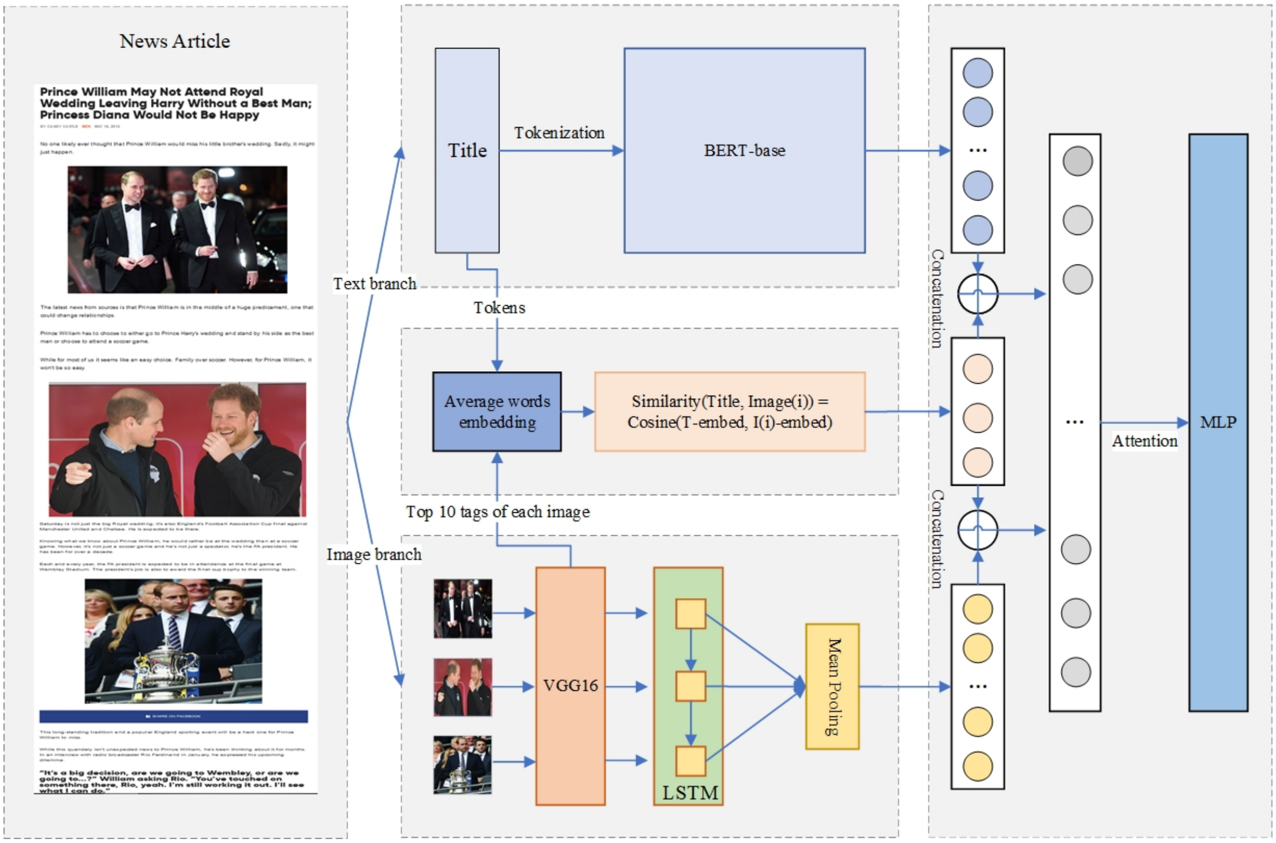
News Article
Prince William May Not Attend Royal Wedding Leaving Harry Without a Best Man; Princess Diana Would Not Be Happy
Text branch
Image branch
Title
Tokenization
BERT-base
Tokens
Average words embedding
Similarity(Title, Image(i)) = Cosine(T-embed, I(i)-embed)
Top 10 tags of each image
VGG16
LSTM
Mean Pooling
Concatenation
Concatenation
Attention
MLP

Fig. 2: Architecture of the multimodal multi-image fake news detection model.

The images in a news post can be seen as a temporal sequence according to image appearance order. To model the sequence information in news post multi-image, we leverage the LSTM model, which has been successfully applied to text classification [27], video classification [29], and other tasks. Generally, LSTM recursively maps the input representations at the current time step to output labels via a sequence of hidden states. The learning process of LSTM is in a sequential manner. Finally, we obtain a hidden state vector at each time step from the last layer of the LSTM.

More formally, the image sequence can be represented as $(I_1, I_2, ..., I_n)$, where n is the number of images in a post. We use the VGG16 to extract the features, that lead to a spatial feature sequence defined as $(X_1, X_2, ..., X_n)$. LSTM networks operate on image VGG16 activations as well as integrate information over multi-image temporal order. We use the LSTM to compute the hidden vector sequence defined as $(h_1, h_2, ..., h_n)$. Finally, the sequence of the hidden outputs are passed into a mean pooling layer over the time steps to produce a single temporal component.

- *Text and image similarity*: To calculate the text-image similarity, we extract the top ten image tags using the pre-trained VGG-16 model. Figure 3 shows two examples of images and the extracted tags. The examples show photos taken during a football match.

After we extract the image tags with VGG-16, every image is represented with 10 image tags. For each tag and text word, we used the word2vec embeddings [16] to estimate the 300-dimension vector by averaging the embeddings. Then the similarity is computed between tags embedding and text embedding using the cosine similarity. The similarity feature is represented by a 3-dimensional vector. More formally, the similarity is calculated using the following equation:

$$Simil(title, image(i)) = cosine(title_{emb}, image(i)_{emb})$$

where $title_{emb}$ refers to the title embeddings and $image(i)_{emb}$ to each of the image tags embeddings.

Finally, these representations were concatenated and fused with attention mechanism. We concatenate the textual (T_f), the visual (V_f) and the semantic component (S_f) and then used as an input to the softmax layer. The softmax function is applied for the output layer to derive a probability representation for each feature. The attention mechanism is applied by multiplying concatenated features with a soft mask of values between zero and one. More formally,

$$F = [T_f, V_f, S_f]$$

$$W = softmax(F \bullet w + b)$$

$$attention = W \bigodot F$$

where w refers to the softmax weight, b to the softmax bias, $\bigodot$ is element-wise multiplication, W represents the feature weights and F refers to the concatenated features.

Finally, we use the Multi-layer Perceptron (MLP) classifier for the final prediction of the article as fake or not.

IV. Experimental Settings

In this section we describe the collection and the experimental settings used to run our experiments.

A. Collection

To run our experiments we use part of the FakeNewsNet [19] collection. In particular, we use the GossipCop posts that are part of the FakeNewsNet collection. GossipCop refers to news that are about celebrities and entertainment. FakeNewsNet contains 5,323 fake and 16,817 real news posted in GossipCop. In the 16,817 real news, there are 8,667 real news articles that have at least one image. In total there are 39,092 images in the real news articles. In addition there are 3,946 news articles that just have one image and 944 news articles with at least 10 images.

In the 5,323 fake news, there are 2,745 real news articles that have at least one image. In total there are 10,899 images in 2,745 news articles. There are 1,357 fake news articles containing only one image and 226 news articles containing at least 10 images.

Due to the imbalance between the classes, we decided to use under-sampling and we randomly selected 5,323 real news posts. After cleaning out the logo and icon images, we managed to collect 2,745 fake news posts and 2,714 real news posts that contain at least one image.

In the crawled news images, there is a large number of non-news content images (such as logos, other news link images, advertisements, icons). Non-news content images are removed through image link deduplication and manual review.

B. Experimental Settings and Hyperparameter Tuning

For the image component of the model, all the images are resized to 224x224x3. Resized images are then passed to VGG-16 and a vector of length 1000 is extracted. We next pass the VGG-16 output to an LSTM layer with 200 hidden units. Finally, the sequence of hidden outputs are passed to mean pooling layer over the time steps to produce a single temporal component.

We use 20% of the collection for test, 10% for validation and the rest 70% for training. We used Keras[4] to build the neural network and VGG16. Table I shows the parameters that are used in the system.

Figure 4 shows the accuracy on training and validation sets for the different values of epochs. We see that the best value was obtained for 60 epochs. Figure 5 shows the F1-metric for the different values of the LSTM hidden layer on the validation set. From the resutls we observe that the best performance was achieved when the number of layers was set to 200.

TABLE I: Neural network parameters.

BERT text sequence length	64
LSTM hidden layer neuron number	200
Fully connected layer	3
Neuron number of each fully connected layer	400, 100, 2
Dropout	0.2
Learning rate	0.00005
Batch size	32
Epochs	60
Optimizer	adam
Training strategy	Early stopping

C. Baselines

To validate the effectiveness of the proposed model, we choose baselines to compare our performance results. We compare the results with the following baselines:

- BERT [7]: This baseline is based only on the textual representation. In particular, it is based on the pretrained BERT-Base available on TensorFlow Hub[5]. The padded and tokenized text is passed into the BERT model to receive word vectors of dimension 768.
- EANN [26]: The EANN model consists of three components: multimodal feature extractor, fake news detector and event discriminator. It is possible to detect fake news using only the multi-modal feature extractor and the fake news detector. Thus, we design a variant of the proposed model, named *EANN-var* which does not include the event discriminator in order to use it as a baseline.
- SpotFake [21]: This is multimodal system that is based on text and image features learned with BERT and VGG-19 pre-trained on ImageNet dataset respectively.

We report F1-metric for the evaluation of the multimodal multi-image system. We use the McNemar test [15] to measure the statistical difference, which is appropriate for comparisons of nominal data.

V. Results

In this section, we present the results of the experiments. First, we show the results of our system when only one image is used. Then, we also present the results of the multimodal multi-image system.

Table II shows the performance results of the experiments on the collection with regards to F1-metric for the three baselines and for the multimodal system when only one image is used. From the results we observe that the baseline that is based on using only the textual representation (BERT) obtains the highest performance. In particular, it achieves a performance of 0.76 with regards to F1-metric. Among the three baselines (i.e., *BERT*, *SpotFake* and *EANN-var*) *BERT* achieves the highest performance and outperforms *SpotFake* by 1.2% and *EANN-var* by 42%. Also, we observe that between *SpotFake* and *EANN-var*, *SpotFake* performs better.

From the results, we also observe that using only features from one image alone achieves a very low performance. However, when this is combined with the textual

[4] https://keras.io/

[5] https://www.tensorflow.org/hub

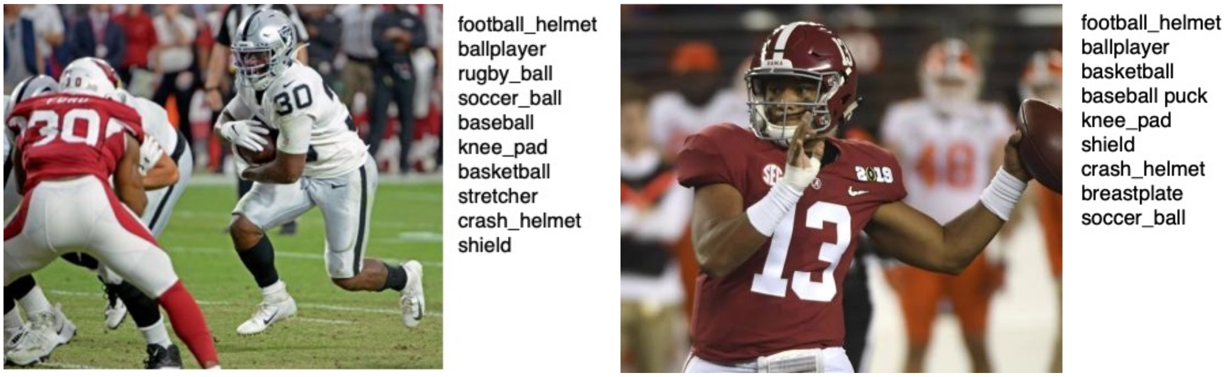

Fig. 3: Examples of images and image tags.

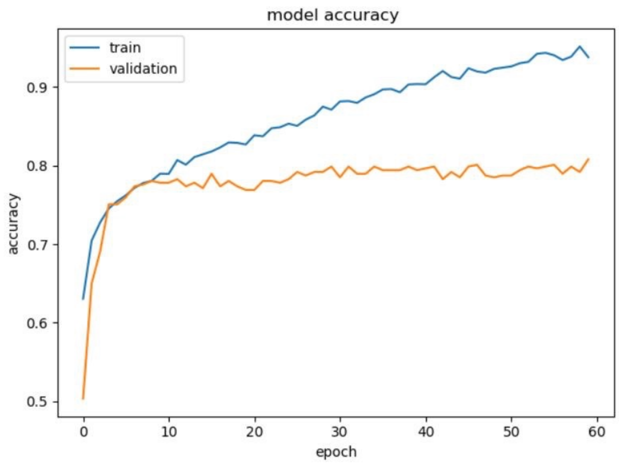

Fig. 4: Accuracy for the different values of epoch.

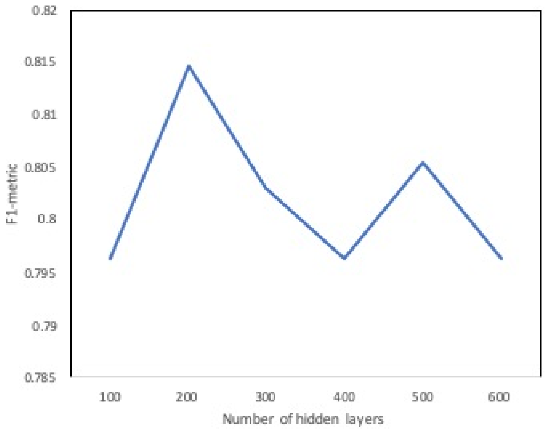

Fig. 5: F1-metric for the different values of LSTM hidden layers.

representation (*1-image-vgg16+BERT+fusion(attention)*), the performance improves. In this case, the system performs similar to the BERT baseline but manages to outperform *SpotFake* and *EANN-var*. In addition, we observe that the attention layer was not helpful in the combination of the textual and visual information extracted from one image (*1-image-vgg16+BERT+fusion*) since the system without attention outperforms the system with the attention by 1.89%. Also, the version without the attention achieves a higher performance compared to *BERT*. The specific combination (1-image-vgg16+BERT+fusion(concatenation)) outperforms BERT by 2.61%. This shows the importance of the visual features in detecting fake news.

Table II also shows that the addition of the image-text similarity was useful in the case of the one image multimodal system (*1-image-vgg16+BERT+similarity+fusion(attention)*). In particular, this combination outperformed all the three baselines and the improvements were statistical significant. Also, the (*1-image-vgg16+BERT+similarity+fusion(attention)*) statistically improves the performance compared to the system without the text-image similarity (*1-image-vgg16+BERT+fusion(attention)*).

Table III shows the results of the experiments of the multimodal multi-image system that combines visual features extracted of three images with textual and semantic representation. Also, the table shows the results of other combinations of the system that we have evaluated. In particular, we explored the performance of extracting visual features from more than three images, using the image-text similarity or not and using the attention layer or not.

From the results we observe that a system that uses only visual features from 3 images (*3-image-vgg16-LSTM*) achieves a higher performance compared to the one that uses only features from only one image (*1-image-vgg16*). This shows the importance of combining the visual features extracted from multiple images of the article. In particular, it achieves an improvement of 58.10% compared to *1-image-vgg16*. Also, we tried to combine visual features from 4 and 5 images. From the results, we observe that they obtain slightly worse results compared to *3-image-vgg16-LSTM*.

In case of combining visual features from three images, we observe that all the different combinations of the multimodal multi-image system outperform the three baselines. In particular, the combination of textual and visual features (*3-*

TABLE II: Performance results of the multimodal system using visual features extracted from one image. The $*$ symbol shows the statistical significance of the systems compared to the *1-image-vgg16+BERT+similarity+fusion(attention)* system.

	F1-score
BERT	0.7628$*$
SpotFake	0.7537$*$
EANN-var	0.4979$*$
1-image-vgg16	0.3678$*$
1-image-vgg16+BERT+fusion(attention)	0.7683$*$
1-image-vgg16+BERT+fusion(concatenation)	0.7830$*$
1-image-vgg16+BERT+similarity+fusion(attention)	0.7683

image-vgg16-LSTM+BERT+fusion(attention)) performs better than *SpotFake*, *EANN-var* and *BERT* and achieves an improvement of 3.33% and 44.05% 2.12% respectively. In the cases of *SpotFake*, *EANN-var* the difference is statistical significant. In the case of the textual and visual combination before incorporating the text-image similarity (*3-image-vgg16-LSTM+BERT+fusion*) we observe that the addition of the attention layer does not impact the performance (*3-image-vgg16-LSTM+BERT+fusion(attention)*).

Then, we also evaluate how the performance changes with the addition of the text-image similarity. From the results we observe that the addition of the text-image similarity to the system manages to improve the performance. In particular, for the system without the attention layer, this improvement is 1.39% (*3-image-vgg16-LSTM+BERT+fusion(concatenation)* compared to *3-image-vgg16-LSTM+BERT+similarity+fusion(concatenation)*) but is not statistical significant, whereas for the system with the attention layer the addition of the text-image similarity manages a statistical significant improvement of 2.049% (*3-image-vgg16-LSTM+BERT+fusion(attention)* compared to *3-image-vgg16-LSTM+BERT+similarity+fusion(attention)*).

Also, we observe that including an attention layer to the multimodal multi-image system improves the performance. In particular, the system that uses attention (*3-image-vgg16-LSTM+BERT+similarity+fusion(attention)*) outperforms the version without attention layer *3-image-vgg16-LSTM+BERT+similarity+fusion(concatenation)* by 0.892%, however the difference is not statistically significant. Finally, the multimodal multi-image system *3-image-vgg16-LSTM+BERT+similarity+fusion(attention)* statistically improves the *BERT* and *SpotFake* baselines by 4.19% and 5.39% respectively.

VI. Conclusion

In this paper we focused on the problem of fake news detection and we proposed a multimodal multi-image system. The proposed system combines textual, visual and semantic information. For the textual representation, we used BERT-Base to better capture the underlying semantic and contextual meaning of the text. For the visual representation, we extracted image tags from multiple images that the articles contained using the VGG-16 model. The semantic information is represented by the image-text similarity that is calculated using the cosine similarity of the title and image tags embeddings. The different components are then concatenated to make the final prediction. Our multimodal multi-image system manages to achieve statistically better results compared to SpotFake, EANN-var and BERT baselines and improves them by 5.39%, 46.02% and 4.19% respectively. Our results showed that combining features from the different components is effective for the fake news detection task and that combining features from multiple images is more effective than using visual features only from one image.

Acknowledgments.: Anastasia Giachanou is supported by the SNSF Early Postdoc Mobility grant under the project Early Fake News Detection on Social Media, Switzerland (P2TIP2_181441). Guobiao Zhang is funded by China Scholarship Council (CSC) from the Ministry of Education of P.R. China. The work of Paolo Rosso is partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31).

References

[1] O. Ajao, D. Bhowmik, and S. Zargari. Sentiment Aware Fake News Detection on Online Social Networks. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2507–2511, 2019.

[2] M. T. Bastos and D. Mercea. The Brexit Botnet and User-Generated Hyperpartisan News. *Social Science Computer Review*, 37(1):38–54, 2019.

[3] A. Bovet and H. A. Makse. Influence of Fake News in Twitter during the 2016 US Presidential Election. *Nature communications*, 10(1):1–14, 2019.

[4] S. Busari and B. Adebayo. Nigeria Records Chloroquine Poisoning after Trump Endorses it for Coronavirus Treatment. CNN, 2020.

[5] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, 2011.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A Large-scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '09, pages 248–255, 2009.

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] D. I. H. Farías, V. Patti, and P. Rosso. Irony Detection in Twitter: The Role of Affective Content. *ACM Transactions on Internet Technology (TOIT)*, 16(3):1–24, 2016.

[9] B. Ghanem, P. Rosso, and F. Rangel. An Emotional Analysis of False Information in Social Media and News Articles. *ACM Transactions on Internet Technology (TOIT).*, 20(2):1–18, 2020.

[10] A. Giachanou, J. Gonzalo, and F. Crestani. Propagating Sentiment Signals for Estimating Reputation Polarity. *Information Processing & Management*, 56(6):102079, 2019.

TABLE III: Performance results of the multimodal multi-image system using visual features from multiple images. The † symbol shows the statistical significance of the systems compared to the *3-image-vgg16-LSTM+BERT+similarity+fusion(concatenation)*. The ∗ symbol shows the statistical significance of the systems compared to the *3-image-vgg16-LSTM+BERT+similarity+fusion(attention)* system.

	F1-score
BERT	0.7628∗ †
SpotFake	0.7537∗ †
EANN-var	0.4979∗ †
1-image-vgg16	0.3678∗ †
3-image-vgg16-LSTM	0.6690∗ †
4-image-vgg16-LSTM	0.6656∗ †
5-image-vgg16-LSTM	0.6465∗ †
1-image-vgg16+BERT+fusion(attention)	0.7683∗
3-image-vgg16-LSTM+BERT+fusion(attention)	0.7792∗
3-image-vgg16-LSTM+BERT+fusion(concatenation)	0.7775
3-image-vgg16-LSTM+BERT+similarity+fusion(attention)	**0.7955**
3-image-vgg16-LSTM+BERT+similarity+fusion(concatenation)	0.7884

[11] A. Giachanou, E. Ríssola, B. Ghanem, F. Crestani, and P. Rosso. The Role of Personality and Linguistic Patterns in Discriminating between Fake News Spreaders. In *Proceedings of the 25th International Conference on Natural Language & Information Systems*, NLDB '20, 2020.

[12] A. Giachanou, P. Rosso, and F. Crestani. Leveraging Emotional Signals for Credibility Detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, pages 877–880, 2019.

[13] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia*, 19(3):598–608, 2017.

[14] D. Khattar, J. S. Goud, M. Gupta, and V. Varma. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *Proceedings of the 2019 World Wide Web Conference*, WWW '19, pages 2915–2921, 2019.

[15] Q. McNemar. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947.

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, NIPS, pages 3111–3119, 2013.

[17] K. Popat, S. Mukherjee, A. Yates, and G. Weikum. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, EMNLP '18, pages 22–32, 2018.

[18] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP '17, pages 2931–2937, 2017.

[19] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fake-NewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *arXiv:1809.01286*, 2018.

[20] K. Shu, S. Wang, and H. Liu. Understanding User Profiles on Social Media for Fake News Detection. In *Proceedings of the 2018 IEEE Conference on Multimedia Information Processing and Retrieval*, MIPR '18, pages 430–435, 2018.

[21] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. SpotFake: A Multi-modal Framework for Fake News Detection. In *Proceedings of the IEEE 5th International Conference on Multimedia Big Data*, BigMM, pages 39–47, 2019.

[22] S.H. Spencer. False Claims of Nationwide Lockdown for COVID-19. Factcheck, 2020.

[23] N. Vo and K. Lee. Learning from Fact-checkers: Analysis and Generation of Fact-checking Language. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '19, pages 335–344, 2019.

[24] S. Vosoughi, D. Roy, and S. Aral. The Spread of True and False News Online. *Science*, 359(6380):1146–1151, 2018.

[25] W. Y. Wang. Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL '17, pages 422–426, 2017.

[26] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao. 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*, KDD'18, pages 849–857, 2018.

[27] M. Yang, W. Tu, J. Wang, F. Xu, and X. Chen. Attention-Based LSTM for Target-Dependent Sentiment Classification. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*, AAAI-17, 2017.

[28] R. T. Yla and W. P. James. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[29] H. Yoo, J. Large-scale Video Classification guided by Batch Normalized LSTM Translator. *CoRR*, abs/1707.04045, 2017.

[30] X. Zhou, J. Wu, and R. Zafarani. SAFE: Similarity-aware Multi-modal Fake News Detection. *arXiv preprint arXiv:2003.04981*, 2020.

[31] D. Zlatkova, P. Nakov, and I Koychev. Fact-checking meets Fauxtography: Verifying Claims about Images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, EMNLP-IJCNLP '19, pages 2099–2108, 2019.

[32] P. Zou and S. Yang. Multimodal Tweet Sentiment Classification Algorithm Based on Attention Mechanism. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 68–79. Springer, 2018.